Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Brian Conroy Date

Spatial Latent Process Models to Overcome Preferential Sampling in Disease Surveillance Systems

By

Brian Conroy
Doctor of Philosophy

Biostatistics and Bioinformatics

Lance Waller, Ph.D.
Advisor

Howard Chang, Ph.D.
Committee Member

Benjamin Lopman, Ph.D.
Committee Member

Robert Lyles, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Spatial Latent Process Models to Overcome Preferential Sampling in Disease Surveillance Systems

By

Brian Conroy
B.S., Harvey Mudd College, CA, 2009
M.S., Emory University, GA, 2018

Advisor: Lance Waller, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2019

Abstract

Spatial Latent Process Models to Overcome Preferential Sampling in Disease Surveillance Systems
By Brian Conroy

Disease surveillance systems are crucial to monitor and predict outbreaks, epidemics and pandemics, as well as to understand the dynamics and trends of diseases over space and time. These systems increasingly rely on complex data collection mechanisms which present particular challenges to the statistician, entailing sampling processes which often violate key assumptions of standard statistical methods. One such mechanism is known as preferential sampling, referring to a stochastic dependency between a spatial process and the locations at which it is observed. While this sampling strategy can lead to considerably biased spatial predictions, few solutions to confront preferential sampling have been proposed in the realm of disease surveillance, despite this potentially deleterious impact. In the first chapter, we propose a novel shared latent process model to correct for preferential sampling in disease surveillance applications, and show by simulation that the practical benefits of such development are reduced bias in parameter estimates and greater accuracy of the estimated disease risk surface. We apply the model to a disease surveillance dataset targeting plague in the rodent population of California, obtaining a substantially improved risk map in comparison to benchmark approaches. In the second chapter, we develop a new multivariate geostatistical model which corrects for preferential sampling when estimating the risk surfaces of a disease common across multiple host species, one which improves spatial predictions by sharing information between species in a hierarchical modeling framework. In the final chapter, we address the dearth of methods to correct for preferential sampling in temporally referenced data by developing a spatiotemporal preferential sampling model, capable of capturing important temporal trends in underlying the disease and sampling processes, yielding more accurate disease risk maps as a result.

Spatial Latent Process Models to Overcome Preferential Sampling in Disease Surveillance Systems

By

Brian Conroy
B.S., Harvey Mudd College, CA, 2009
M.S., Emory University, GA, 2018

Advisor: Lance Waller, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2019

Acknowledgments

This dissertation presented several challenges, statistical, and computational, and mental. If I have made any progress on any of these fronts it comes with deep indebtedness to my advisor Dr. Lance Waller, especially for his wealth of insights, expertise in spatial statistics, and continued encouragement to produce translatable results. In addition, the methodological developments of my dissertation have benefitted greatly from the helpful feedback provided by my committee members Dr. Howard Change, Dr. Benjamin Lopman, and Dr. Robert Lyles. But the development of these methods means nothing without application. For that I am exceedingly grateful to the California Department of Public Health, in particular the invaluable collaboration with Dr. Greg Hacker, Dr. Mark Novak, and Dr. James Tucker. Lastly, I have received tremendous help from Dr. Ian Buller, through his expertise in the mechanics of the plague surveillance system, as well as the vast amount of work he undertook in preparing and managing the data which inspired my analyses. Thank you all!

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Preferential Sampling in Disease Surveillance

1.1 Introduction

Disease surveillance systems are crucial to monitor and predict outbreaks, epidemics and pandemics, as well as to understand the dynamics and trends of diseases over space and time. These systems increasingly rely on complex or unconventional data collection mechanisms (Choi et al., 2016; Plowright et al., 2019) which present particular challenges to the statistician, entailing sampling processes which often violate key assumptions of standard statistical methods. Several types of sampling issues may characterize disease surveillance applications, notably opportunistic sampling (Zimmer and Lee, 2019), detection error (Tabak et al., 2019), and preferential sampling (Cecconi et al., 2016). Failing to accurately adjust for these influences in statistical modeling has the potential to greatly distort parameter estimates and predicted risk surfaces, the practical consequences of which are, at best, an erroneous understanding

of the relationship between key covariates and disease risk, and at worst, a mischaracterization of areas at risk, thereby harming the ability of the system to optimally monitor and respond to the disease.

Project 1 considers a particularly challenging application of disease surveillance: the surveillance of the zoonotic hosts of plague in the American southwest. This application motivates the development of novel methods to correct for preferential sampling, a sampling mechanism in which observation sites tend to be assigned more frequently to areas at high risk for the disease. We propose a shared latent process model to address this sampling strategy, and show by simulation that the practical benefits of such development are reduced bias in parameter estimates and greater quality of the estimated disease risk surface.

While the primary methodological focus of Project 1 involves novel solutions to correct for preferential sampling in the context of disease surveillance, preferential sampling itself arises in several other domains. More generally, preferential sampling is a data collection strategy in which there is a stochastic dependence between the locations at which a spatial process is observed and the value of the process itself. Applications which often rely on preferentially sampled data include air pollution monitoring (Lee et al., 2011), mineral exploration (Veneziano and Kitanidis, 1982), species distribution modeling (Gelfand and Shirota, 2019), and real estate pricing (Paci et al., 2019). Preferential sampling may arise for a variety of reasons. In the context of disease surveillance, one rationale is to collect a maximum volume of information from areas of particular concern given the constraint of limited resources, thereby allowing for a rapid response to outbreaks by monitoring areas at greatest risk. Alternately, when conducting surveillance on a rare disease, it may be impractical to employ random sampling, given the low probability that each sample has of containing any disease positive results. Or, when monitoring a zoonotic disease, the best use of

limited resources may be to surveil areas of high consequence, where emergence of the disease would come at a particularly high human or economic cost, in contrast to areas where the disease may reside in the host species but where transmission to humans is unlikely. A similar motivation appears in species distribution modeling, where random sampling is typically too impractical or unlikely to observe the species of interest to be useful (Gelfand and Shirota, 2019). In the context of air pollution monitoring, a network of observation sites may be specifically designed to measure extreme values in order to identify noncompliance with pollution regulations (Chang et al., 2007). These concerns are largely separate from obtaining an unbiased estimate of the spatial process of interest over the entirety of the study region. In point of fact, the downside of preferential sampling is often a negative impact on the quality of parameter estimates and spatial predictions.

While sampling at high response locations is often the optimal choice from a managerial or budgetary perspective, the drawback is that it may result in a statistically biased estimation of the spatial process of interest (Diggle et al., 2010; Lee et al., 2011; Pati et al., 2011; Gelfand et al., 2012; Lee et al., 2015; Gelfand et al., 2019). This phenomenon is due to the fact that under preferential sampling, the locations of observation sites, X, are stochastically dependent on Y, the observed response, whereas conventional geostatistical methods typically assume X to be fixed, allowing for the joint distribution $[Y, X]$ to be simply factorized as $[Y, X] = [Y][X]$, where $[.]$ denotes distribution. However, preferential sampling may be ignored if the association between the surface of interest and sampling locations can in fact be entirely explained by shared spatial covariates, for which one adjusts (Pacifici et al., 2016). Otherwise, erroneous inferences may result from failing to account for the stochastic dependence between X and Y. Ultimately, when considering preferential sampling in disease surveillance, there are two contrasting aims which conflict here: 1) to monitor most sensitive areas in order to respond quickly to high impact threats, all the while

using constrained resources, and 2) to estimate an unbiased risk surface over a broad extent. Preferential sampling lends itself to the former, while possibly impacting the ability to perform the latter. Thus our goal is to propose a model that can estimate a statistically unbiased risk surface from preferentially sampled data.

The geostatistical literature contains a rich body of work for addressing preferential sampling. Essential to these methods is the strategy of building a joint model for locations and the measured response, wherein the locations are treated as random quantities, typically realizations of a point process whose intensity is itself related to the process of interest. Most notable among this body of work is the shared latent process model of Diggle, Menezes and Su (2010), which inspired several other works (Pati et al., 2011; Lee et al., 2011; Lee et al., 2015; Pennino et al., 2019). While the shared latent process model, or variants thereof, have inspired a substantial body of preferential sampling analyses in the realm of environmental pollutant monitoring, analyses of preferentially sampled disease surveillance data seldom have attempted to correct for the sampling process. The few exceptions to this trend arise in the area of veterinary health monitoring (Rinaldi et al., 2015; Cecconi et al., 2016). To adjust for preferential sampling in the analysis of parisitological livestock risk, Cecconi et al. (2016) estimate spatially varying sampling probabilities from the fraction of sampled farms within each grid cell in a discretization of the study region. These probabilities are subsequently included as covariates in a geostatistical model of disease risks. While preferential sampling adjustments have proven successful in veterinary health surveillance, these applications hinge upon the ability to calculate spatially varying sampling probabilities, given by the fraction of sampled farms in a region, which entails knowing the total number of farms in a given area as the denominator. Our methodological developments in this project focus on a scenario more common in zoonotic disease surveillance, wherein the denominator of any particular grid cell is unknown.

The methods introduced in Project 1 are grounded in the application of plague surveillance in the American southwest. Plague is an infectious disease caused by the Gram-negative bacterium *Yersinia pestis*, typically transmitted through the bite of an infected flea, although other forms of transmission occur, including droplet contact, airborne transmission, or direct or indirect contact with an infected individual (Eisen et al., 2006). Three main forms of plague exist, namely pneumonic, septicemic, and bubonic, the last of which is infamous as the causative agent of the "Black Death" in medieval Europe, estimated to have killed over 20 million people (Byrne 2004). However, there have been in fact 3 historical outbreaks of bubonic plague reaching pandemic levels: the plague of Justinian, affecting the Middle East and Mediterranean in the 6th century CE, the Black Death beginning in 1347, spreading from western Asia throughout the Middle East, Mediterranean, and Europe, as well as the 1894 bubonic plague originating in Canton and Hong Kong, which subsequently extended throughout most of Asia and India, spreading via international shipping to ports as distant as San Francisco and Glasgow (Benedict, 1996). While there has not been a plague pandemic since the early 20th century, the disease is still prevalent in over 20 countries (Chanteau et al., 2000), and thus remains a nontrivial public health concern.

However, due to the present rarity of plague in the human population of the southwestern United States, it is not practical to surveil humans alone for the disease in this region. A much better target of surveillance consists of the zoonotic hosts of plague, where prevalence persists at roughly 10% in the coyote (*Canis latrans*) population and between 2-5% in the rodent population. Monitoring plague in its zoonotic hosts is especially important given that many plague outbreaks in the human population are preceded by epizootics, or outbreaks among animals (Migliani et al., 2006). The specific plague surveillance system we consider is operated by the California Department of Public Health, and monitors plague in the rodent family of squirrels, known

as *Sciuridae* or *Sciurids*, comprising a total of 21 different species. The surveillance system collects data by conducting a series of sampling events at locations throughout California. At each sampling location, Sciurids are trapped and subsequently tested for *Yersinia pestis*, primarily through F1 antigen tests. The surveillance system predominantly assigns sampling locations to high risk or high impact areas, where risk is assessed to be high in what are viewed as plague endemic regions, as determined by historic cases of plague in humans or recovered Sciurid specimen, and high impact areas are regions where cases of plague in humans would be particularly damaging, such as in national parks. In this sense the dataset is preferentially sampled. The key challenge here is thus to correct for the effects of preferential sampling in order to obtain improved parameter estimates and predicted risk surfaces.

In Project 1 we propose a shared latent process model to adjust for preferential sampling when estimating disease risk maps. The model includes two main components, one locational, describing the distribution of observation sites, and the other describing the spatial abundances of observed disease positive and negative specimen. The locational component models the distribution of observation sites as an inhomogeneous point process, whose intensity function is formulated in terms of a latent spatial process. This process is shared with the intensity function of the disease component, which describes the distribution of cases and controls also in terms of an inhomogeneous point process. The model we propose in Project 1 is temporally aggregated in the sense that it pools data collected by the surveillance system between 1983 and 2015. The intent here is to develop and assess a baseline model to adjust for preferential sampling, before considering temporal trends in the third project of this dissertation.

The following section presents the statistical background underpinning the developments of Project 1, namely point processes, spatial epidemiology, Gaussian processes,

spatial process models, preferential sampling models, Hamiltonian Monte Carlo, and spatial downscaling. We then describe the novel methods introduced by Project 1, adapting existing preferential sampling methods to the application of disease surveillance. The methods section is then followed by a series of simulation studies, demonstrating practical benefits of these developments in the form of reduced bias in the estimates of parameters of scientific interest, as well as improved disease risk surface estimation compared to benchmark approaches which do not address the sampling process.

1.2 Methods

1.2.1 Introduction

The basis of geostatistics is to make inferences about some spatial surface from samples drawn over a limited set of locations. In the context of disease surveillance, our target is the risk surface of the disease. Traditional geostatistical models assume that the pattern of sampling locations is not related to the spatial process being measured. That is, samples are thought to be statistically independent from the response of interest. A problem arises in that uniform random sampling is impractical in many real world scenarios, especially those pertaining to disease surveillance. Instead, budgetary constraints and other considerations may dictate a pattern of sampling which is in fact related to the spatial process of interest, wherein samples are collected in areas of high value for the spatial surface of interest. We refer to this sampling scheme as *preferential sampling*. Methods to correct for preferential sampling typically rely on modeling the distribution of observation sites in the framework of point processes, which we now examine.

1.2.2 Point Processes

Studies concerned with the spatial distribution of disease risk typically confront outcomes which are either point referenced, corresponding to a discrete set of locations at which events are recorded (such as the presentation of a diseased individual), or areal, that is, aggregated in space, such as the total number of diseased individuals within a county or other spatial unit (Waller and Gotway, 2004). For our purposes we first turn our attention to a brief overview of spatial point processes, a key framework on which our developments hinge. Point processes describe random patterns of observed events in space, with obvious relevance to several natural phenomena across disciplines, such as the positions of trees in a field, the home locations of infected patients, the position of stars in outer space, or the locations of human developments in some region, to name a few. The statistical treatment of point processes begins with a description of the first-order properties of the process, by way of the intensity function. The intensity function of a point process describes the expected abundance of points occurring in any subregion of space.

Definition 1. *For location x and $N(B)$ denoting the number of points in any subregion B, the intensity function of a univariate spatial point process is defined as*

$$\lambda(x) = \lim_{|dx| \to 0} \left(\frac{E[N(dx)]}{|dx|} \right)$$

From this definition it follows that the expected number of points in any subregion B is given by $N(B) = \int_B \lambda(x)dx$. The second-order properties are specified by the second-order intensity function.

Definition 2. *For locations x and y, with $N(B)$ denoting the number of points in*

any subregion B, the second-order intensity function is defined as

$$\lambda_2(x,y) = \lim_{|dx|,|dy|\to 0} \left(\frac{E[N(dx)N(dy)]}{|dx||dy|} \right)$$

We now consider two prominent types of point processes: binomial and Poisson processes.

Definition 3. *Let f be a density function on set $B \subseteq S$, let $n \in \mathbb{N}$. A point process X consisting of n i.i.d. points with density f is a binomial point process of n points in B with density f, written $X \sim Binomial(B, n, f)$.*

As a simple example, the binomial point processes specified for distribution function $f = 1/|B|$ describes a uniform scattering of n points over the study region. The next class of point processes, Poisson processes, are defined with respect to binomial processes.

Definition 4. *A point process X on S is a Poisson point process with intensity function λ if: 1) for any $B \subseteq S$ with $\mu(B) < \infty$, $N(B) \sim Poisson(\mu(B))$, where $\mu(B) = \int_B \lambda(s)ds$. 2) for any $n \in \mathbb{N}$ and $B \subseteq S$ with $0 < \mu(B) < \infty$, conditional on $N(B) = n, X_B \sim Binomial(B, n, f)$ with $f(x) = \lambda(x)/\mu(B)$. We then write $X \sim Poisson(\lambda)$.*

If λ is constant, X is called a *homogeneous point process* with rate λ, otherwise it is said to be an *inhomogeneous point process* (Figure 1.1).

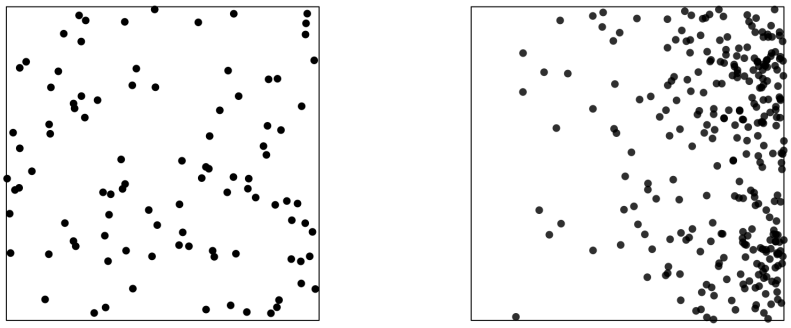

Figure 1.1: Homogeneous (left) and inhomogeneous (right) point processes. For the left, $\lambda = 100$, and for the right, $\lambda(x, y) = 100\exp(5x)$.

Application to Spatial Epidemiology

The point process framework readily lends itself to the analysis of spatial variation in risk, as explained by Diggle (2013), who conceptualizes this application as follows. First let $r(x)$ be the probability an individual at location $x \in \mathbb{R}^2$ will be a case. The intensity functions of cases and controls are, respectively,

$$\lambda_1(x) = r(x)\lambda(x)$$

$$\lambda_2(x) = c(1 - r(x))\lambda(x)$$

where $\lambda(x)$ is the intensity of the underlying population and c is a constant. Consequently, the disease odds $r(x)/(1 - r(x))$ can be calculated up to a constant as

$$c^{-1}\frac{r(x)}{1 - r(x)} = \lambda_1(x)/\lambda_2(x)$$

We then solve the above equation for $r(x)$ and plug in estimates of $\lambda_1(x)$ and $\lambda_2(x)$ to estimate the disease risk surface. A great number of studies have used point processes

for risk surface estimation in this fashion, including Benes et al. (2005) with tick-born encephalitis, Diggle et al. (2005) for nonspecific gastroenteric disease in the UK, Liang et al. (2009) jointly modeling point patterns of different cancer types, Ahn et al. (2014) modeling case patterns of diarrheal disease within communities spread over a river network in northwestern Ecuador, or Junior et al. (2015), who incorporated spatially varying effects of individual covariates into the intensity function of a point process, applied to the pattern of cerebrovascular deaths in Rio de Janeiro. Point processes are thus useful tools for the field of spatial epidemiology, with the provision of certain key caveats. Firstly, as noted by Diggle (2013), confining a specimen to a particular point in space fails to account for movement, complicating analysis by the fact that the place in which an individual is recorded may not be representative of where disease exposure occurred. Another issue, known as the ecological fallacy, arises when relevant covariates are only available at the aggregate level, such as by county or other geographic unit, rather than at the point level, in which case an unbiased estimation of the relationship between disease risk and covariates may not be recoverable (Wakefield and Shaddick, 2006).

1.2.3 Gaussian Processes

We have seen in the previous section that a key consideration of spatial epidemiology concerns the possibility of spatial variation in risk, and what known factors, if any, influence this variation. From the perspective of point process modeling, spatial variation in risk implies inhomogeneity in the ratio of case and control intensity functions, $\lambda_1(x)/\lambda_2(x)$. The key to modeling the point patterns of cases and controls then becomes how to construct the intensity functions, $\lambda_1(x)$ and $\lambda_2(x)$. Intensity functions are commonly modeled to be log-linear in certain covariates, such as $\log(\lambda(x)) = z(x)^T\beta$, where $z(x)$ are spatial covariates. However, in many cases

the covariates $z(x)$ are inadequate to capture all variation present in the data. When unexplained spatial variation is present, it becomes valuable to introduce spatially structured residuals into the construction of the intensity function to capture this additional variation. Gaussian processes are one widely used solution in this regard, although they have application well beyond spatial statistics to several other regression and modeling problems.

A Gaussian process is a set of random variables such that any finite subset follows a multivariate normal distribution, whose mean and covariance are specified by known functions. That is, supposing $(Y_1, \ldots, Y_n)$ are random variables realized from a Gaussian process, then

$$(Y_1, \ldots, Y_n)^T \sim \text{Normal}(\mu, \Sigma[k(.,.)])$$

where μ is the *mean function* and $\Sigma[k(.,.)]$ is the covariance matrix, which has been constructed by the *covariance function* $k(.,.)$ of the Gaussian process, which specifies the covariance between any two random variables in the process. In the context of spatial statistics we consider subsets of the form $(Y_1(x), \ldots, Y_n(x))$, where each $Y_i(x)$ is taken at some point x in typically 2 or 3 dimensional space. In this case, the mean μ of the set of n arbitrary random variables may itself be a function of x. In addition, the i, j-th element of the covariance matrix is computed by the covariance function, $k(x, x')$, which takes as input the locations of the ith and jth points.

Given that each element of the covariance matrix is calculated from the covariance function $k(x, x')$, it is clear that certain restrictions must be placed on $k(x, x')$ such that the resulting covariance matrix is valid, i.e. symmetric and positive definite. A commonly used covariance function is the exponential function,

$$k(x, x'; \sigma^2, \theta) = \sigma^2 \exp(-||x - x'||/\theta)$$

where $||x - x'||$ is the distance between points x and x', θ the nonnegative *range* parameter controlling the scale of the spatial association, and σ^2 the marginal variance, representing the magnitude of covariance as distance shrinks to zero. Another commonly used covariance function is the squared exponential covariance function, $k(x, x'; \phi, \theta) = \phi \exp(-||x - x'||^2/\theta)$, possessing the property of smoothness, unlike the exponential function. A third popular option is the Matérn covariance function,

$$k(d; \nu, \theta, \sigma^2) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\sqrt{2\nu}\frac{d}{\theta}\Big)^{\nu} K_{\nu}\Big(\sqrt{2\nu}\frac{d}{\theta}\Big)$$

Here, d is the distance between points, K_ν the modified Bessell function, and ν and θ nonnegative parameters, with θ the range parameter and ν controlling smoothness. An important quality of the Matérn and exponential covariance functions is that of stationarity. A covariance function is *stationary* if it depends only on the separation between the two points, rather than the absolute locations of the points in space. In addition, a covariance function is *isotropic* if the distance metric used is Euclidean, and does not depend on direction. Non-isotropic functions are relevant when the directionality between points is of significance, as may happen in scenarios such as monitoring pollutant levels along a river, where the direction of flow influences the correlation in responses between points along the waterway. We concern ourselves in this project with stationary, isotropic covariance functions. In the next section we examine a class of point processes which incorporate Gaussian processes into their intensity functions, before delving into the central problem of this project: how preferential sampling influences the predictions made concerning a Gaussian process at novel locations.

Log Gaussian Cox Processes

In many applications, particularly epidemiological ones (Diggle et al., 2005; Liang et al., 2008; Junior et al., 2014), an intensity function constructed from fixed covariates alone may be inadequate to explain the observed point patterns of cases or controls. It is often useful to introduce a stochastic component to the intensity function, especially when there is unexplained spatial autocorrelation between points. This reasoning motivates Cox processes.

Definition 5. *Suppose that* $Z = \{Z(\xi) : \xi \in S\}$ *is a nonnegative random field so that with probability 1,* $Z(\xi)$ *is locally integrable. If the conditional distribution of* X *given* Z *is a Poisson process on* S *with intensity function* Z*, then* X *is said to be a Cox process driven by* Z*.*

In short, a Cox process is a Poisson process with a stochastic intensity function. The family of log Gaussian Cox Processes (LGCP) is a particular type of Cox process wherein the intensity function is a Gaussian field.

Definition 6. *Suppose that* X *is a Cox process driven by* $Z = exp(Y)$ *where* Y *is a Gaussian field. Then* X *is said to be a log Gaussian Cox process (LGCP).*

LGCPs are widely used in spatial epidemiology (Benes et al., 2005; Ahn et al., 2014), as well as in methods to confront preferential sampling, as we shall see shortly. Further discussion of the theory and application of point processes, particularly LGCPs, has been provided by Møller and Waagepetersen (2003).

1.2.4 Spatial Process Models

In the previous section we have seen how Gaussian processes can be incorporated into the intensity functions of point processes to capture additional spatial variation beyond what covariates alone may describe. Now, we examine another set of models derived from Gaussian process, i.e. spatial process models. The preferential sampling literature focuses on the effects of nonrandom sampling on this class of models. The conventional spatial process model describes a response $Y(s)$ at location s as the sum of a deterministic mean function $\mu(s)$ and residual component $w(s)+\epsilon(s)$, where $w(s)$ and $\epsilon(s)$ are spatial and nonspatial residuals, respectively:

$$Y(s) = \mu(s) + w(s) + \epsilon(s)$$

The deterministic mean $\mu(s)$ is typically a function of spatial covariates, $\mu(s) = x(s)^T\beta$, and the $w(s)$ follow a Gaussian process with mean 0 and stationary covariance function $k(x, x'; \sigma^2, \theta)$, where σ^2 is the marginal variance and θ the spatial range. $w(s)$ is intended to capture additional spatial variation not explained by the mean component $\mu(s)$. Residuals $\epsilon(s)$, are assumed independent with mean zero and variance τ^2, referred to as the *nugget* effect. The nonspatial residuals $\epsilon(s)$ are often interpreted as measurement error or noise accompanying repeat measurements at a particular location, or as micro-scale variability, i.e. variation in the response at distances smaller than the distance between sites observed in the data.

We now proceed to an introduction of spatial kriging, the process by which response values $Y(s)$ of the spatial process model are predicted at new, unobserved locations. Then, in the subsequent section, we examine how preferential sampling deteriorates the quality of kriging predictions, a core problem this dissertation seeks to address.

Kriging

Kriging uses Gaussian field theory to define optimal spatial prediction. We seek to use observations of a spatial process Y observed at locations $(s_1, \ldots, s_n)$ to find an optimal linear predictor for the value of Y at any unknown location s_0 that minimizes the mean squared error. That is, we seek a function of the observed data, $f(Y)$, which minimizes

$$E[(Y(s_0) - f(Y))^2|Y]$$

In their explanation of kriging, Bannerjee et al. (2004) show that this expectation can be rewritten as

$$\begin{aligned} E[(Y(s_0) - f(Y))^2|Y] &= E[(Y(s_0) - f(Y) + E[Y(s_0)|Y] - E[Y(s_0)|Y])^2|Y] \\ &= E[(Y(s_0) - E[Y(s_0)|Y])^2|Y] + (E[Y(s_0)|Y] - f(y))^2 \end{aligned}$$

and since $(E[Y(s_0)|Y] - f(y))^2$ is nonnegative, it follows that $E[(Y(s_0) - f(Y))^2|Y] \geq E[(Y(s_0) - E[Y(s_0)|Y])^2|Y]$. Therefore, the function which minimizes the mean squared error must be the conditional expectation of $Y(s_0)$ given the data, $E[Y(s_0)|Y]$. In order to estimate this conditional expectation we return to the spatial process model given in the previous section,

$$Y(s) = \mu(s) + w(s) + \epsilon(s)$$

which implies that $Y = (Y(s_1), \ldots, Y(s_n) \sim N(X\beta, \sigma^2 H(\theta) + \tau^2 I)$. Here $H(\theta)$ has

i, jth element $\rho(x_i, x_j; \theta)$, where ρ is a valid, positive definite correlation function with range parameter θ. Multivariate normal theory holds that if

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\Big(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \Big)$$

then

$$E[Y_1|Y_2] = \mu_1 + \Omega_{12}\Omega_{22}^{-1}(Y_2 - \mu_2)$$
$$Var[Y_1|Y_2] = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$$

In the context of kriging we let $Y_1 = Y(s_0)$, the value of the process at unobserved location s_0, and $Y_2 = Y = (Y(s_1), \ldots, Y(s_n))$, as well as $\Omega_{11} = \sigma^2 + \tau^2$, $\Omega_{22} = \sigma^2 H(\theta) + \tau^2 I$, and $\Omega_{12} = \gamma^T$ where $\gamma^T = (\sigma^2\rho(d_{01}; \theta), \ldots, \sigma^2\rho(d_{0n}; \theta))$, with the understanding that $\rho(d_{0i}; \theta)$ represents the correlation between the ith observed point and unobserved point s_0. Therefore, it follows from the above conditional expectation and variance formulas that the best linear predictor for $Y(s_0)$, and its variance, are

$$E[Y(s_0)|Y] = x_0^T\beta + \gamma^T\Sigma^{-1}(y - X\beta)$$
$$Var[Y(s_0)|Y] = \sigma^2 + \tau^2 - \gamma^T\Sigma^{-1}\gamma$$

From a Bayesian perspective, kriging approximates the posterior predictive distribution of the response value at an unobserved location, $p(Y(s_0)|Y, X, x_0)$. Collecting model parameters into a single vector as $\phi = (\beta, \sigma^2, \theta, \tau^2)$, the predictive distribution

is given by

$$p(Y(s_0)|Y,X,x_0) = \int p(Y(s_0),\phi|Y,X,x_0)d\phi$$
$$= \int p(Y(s_0),|Y,X,x_0)p(\phi|y,X)d\phi$$

Here, $p(Y(s_0),|Y,X,x_0)$ has the conditional normal distribution provided above. If posterior samples $\phi^{(1)},\ldots,\phi^{(m)}$ are drawn from $p(\phi|y,X)$ by Markov Chain Monte Carlo methods, then the predictive distribution is typically estimated as

$$\hat{p}(Y(s_0),|Y,X,x_0) = G^{-1}\Sigma_{g=1}^{G} p(y_0|Y,\phi^G,x_0)$$

We thus draw kriged samples from the conditional normal distribution where the gth MCMC sample for ϕ is plugged in to the conditional mean and variance formula. If it is of interest to jointly predict the value of the response at multiple locations, then the same approach as that used to krige for a single location may be used, with the understanding that Y_1 in the conditional multivariate normal formula is now taken as a vector of response values.

Kriging and Preferential Sampling: Illustrated Example

Under normal circumstances, the distribution of sample sites s is assumed not to relate to the response. But many real world scenarios, such as mineral exploration, pollution monitoring, and disease surveillance, to name a few, often conduct measurements predominantly in areas which correspond to high values for the process of interest. The crux of the problem we investigate in this project is that, under

such a sampling strategy, kriging typically yields biased predictions (Diggle et al., 2010; Pati et al., 2011; Lee et al., 2011; Gelfand et al., 2012; Lee et al., 2015). We can illustrate the effect with the following example. Suppose we discretize the state of California into a study region containing 405 equally sized square grid cells, and simulate a stationary, mean zero spatial Gaussian process over the centroids of these cells, choosing an exponential covariance function $k(x, x'; \sigma^2, \theta)$, where $\sigma^2 = 12$ and $\theta = 6$. We refer to the simulated spatial random effects as $w = (w_1, \ldots, w_{405})$. We then preferentially sample w by selecting the centroids corresponding to the 50 greatest values of w as observation sites, collecting the values of w at these sites in a vector $W_2 = (w_{(1)}, \ldots, w_{(50)})$. If we let W_1 refer to the vector of random effects not observed in the previous step, then from the conditional expectation of a multivariate normal random variable it follows that

$$E[W_1|W_2] = \Omega_{12}\Omega_{22}^{-1}W_2$$

Here we assume the true values of σ^2 and θ are known and can thus be used directly to calculate Ω_{12} and Ω_{22}. For comparison, we also randomly sample 50 random effects, and use the above formula to calculate the conditional mean of the unobserved w values under random sampling. Taking the conditional means $E[W_1|W_2]$ to be our predictions of the Gaussian process, we can plot true versus predicted W_1 values for both sampling strategies, i.e. preferential and random (Figure 1.2). We see that the predicted values under preferential sampling grossly overestimate the true values, whereas the distribution of error for random sampling is much more even and of lesser magnitude.

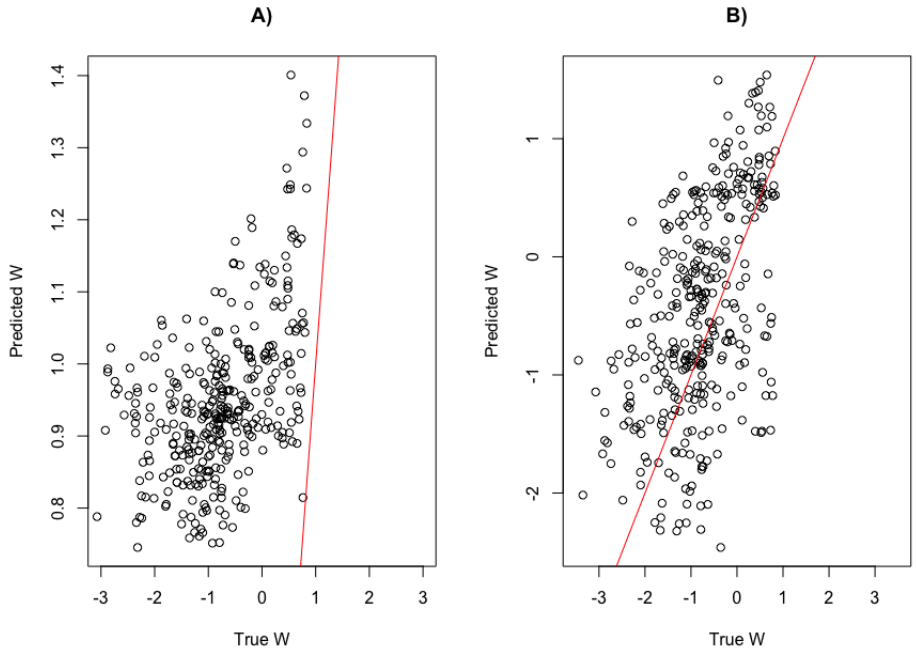

Figure 1.2: Comparison of true versus predicted random effect values under A) preferential sampling and B) random sampling. The red lines are of slope 1, intercept 0.

We now proceed to an overview of the existing solutions in the literature to correct for the bias-inducing effect of preferential sampling, before introducing our novel proposed method.

1.2.5 Preferential Sampling

A seminal contribution to the analysis of preferentially sampled data is the shared latent process model proposed by Diggle, Menezes and Su (2010), hereafter referred to as the SLP model. The SLP model treats observation sites as stochastic, in particular, as realizations of a point process with intensity driven by a generalized spatial process. This spatial process is, in fact, the process of geostatistical interest, thus enabling sites to be assigned to areas of high response value according to some strength of preference. We outline this model as follows. Let S denote the unobserved spatially continuous

process of interest on region A, let X denote a point process on A and let Y be a set of measured values of S at each point of X. The authors then define a model with the following 3 assumptions:

1) S is a stationary mean-zero Gaussian process with some known covariance function $k(.,.)$ 2) Conditional on S, X is a log Gaussian Cox process with intensity $\xi(x) = \exp(\alpha + \beta S(x))$. 3) Conditional on S and X, Y is a set of mutually independent Gaussian variates with $Y_i \sim N(\mu + S(x_i), \tau^2)$. That is,

$$
\begin{aligned}
S &\sim \mathcal{GP}(0, k(,.,)) \\
X|S &\sim \mathcal{LGCP}(\xi(x)) \\
\xi(x) &= \exp(\alpha + \beta S(x)) \\
Y_i|S(x_i) &\sim N(\mu + S(x_i), \tau^2)
\end{aligned}
$$

The SLP model explicitly captures stochastic dependence between S and X through the inclusion of S in the intensity function $\xi(x)$. The parameter β controls the strength of preferential sampling, inducing stronger preference toward high response areas as β increases. Readers familiar with marked point processes (Ho and Stoyan, 2008) will note that the SLP model can be viewed as just that, a marked point process with points described by X and marks given by the measured values Y_i at each point. This structural design, featuring one model component describing the distribution of observation sites and another capturing the values of the response measured at the given sites, is a motif which reoccurs throughout the majority of methods to correct for preferential sampling, including those proposed in this dissertation.

Diggle et al. (2010) then conduct a number of simulations to probe the performance

of traditional geostatistical methods in analyzing preferentially sampled data. Substantial bias in the empirical variogram estimate was found when the data were preferentially sampled, as well as large positive bias in the ordinary kriging estimate. For application, the SLP model was fit via a Monte Carlo maximum likelihood approximation to a data set consisting of heavy metal biomonitoring in Galicia, northern Spain, in which surveys were conducted more extensively in regions where large gradients in lead concentration were expected. Substantial differences in predicted lead concentration surfaces were found between the SLP model and standard geostatistical analysis. This study thus highlights the potential bias in parameter estimates and model predictions which may result when preferential sampling is ignored.

While the model proposed by Diggle et al. (2010) is the progenitor of much of the subsequent research efforts confronting preferential sampling, it is not without drawbacks. As noted in the discussion by Rue at al. (2010), the use of the method in practice may be hindered by the computationally intensive nature of its Monte Carlo likelihood approximation. But even more alarming is the objection recently raised by Dinsdale and Salibian-Barrera (2019), who showed that Diggle's Monte Carlo likelihood may in fact not approximate the desired likelihood function at all under preferential sampling due to previously overlooked implicit conditioning on the distribution of observation sites in formulating the Monte Carlo likelihood. However, the proposed model of this dissertation avoids Monte Carlo likelihood approximations altogether and so cirvumvents this issue.

The methods proposed in Project 1 of this dissertation differ from those of Diggle et al. in other key respects. First, the SLP model has a spatially constant mean μ, whereas we propose the local mean to vary with location through covariates, not just covariation via $S(x)$. Second, the estimated disease risk surface differs from the lead concentration surface in that it is composed of two additional sub-surfaces, that is,

case and control surfaces. This distinction calls for a number of structural changes to be made to the model, in order to capture the separate effects which preferential sampling will have on cases and controls.

A prominent extension of the SLP model comes from Pati et al. (2011), who offer a fully Bayesian version of the model, implemented via MCMC rather than Monte Carlo maximum likelihood, along with other notable structural changes. This model proceeds as follows. Let S be a Gaussian process with mean zero and known covariance function $k(.,.)$, and X the point pattern of observation sites. Then the authors assume that $X|S$ follows a log Gaussian Cox process with intensity $\xi(x)$, where $\xi(x)$ is log-linear in $z(x)\beta_\xi + S(x)$, for spatial covariates $z(x)$. Then for value Y_i measured at site x_i, Y_i conditional on $x_i, \xi(x_i)$ is distributed as $N(\lambda(x_i) + \alpha\log(\xi(x_i)), \sigma^2)$.

$$
\begin{aligned}
&S \sim \mathcal{GP}(0, k(,.,)) \\
&X|S \sim \mathcal{LGCP}(\xi(x)) \\
&\xi(x) = \exp(z(x)\beta_\xi + S(x)) \\
&Y_i|\xi(x_i) \sim N(\lambda(x_i) + \alpha\log(\xi(x_i)), \sigma^2) \\
&\lambda(x) = \exp(z(x)\beta_\lambda + \lambda_r(x)) \\
&\lambda_r(x) \sim \mathcal{GP}(0, k(,.,))
\end{aligned}
$$

Here, unlike the SLP model, the local mean of the response is generalized through construction as $\lambda(x)$, a function of spatial covariates $z(x)$ and Gaussian process λ_r, which captures unexplained spatial variation in measurement values. $\alpha\log(\xi(x_i)$ is included in the response mean to correct for preferential sampling, with α indicating the strength of preferential sampling. For instance, $\alpha > 0$ corresponds to observation

sites being placed preferentially in regions of high response value.

Additional adaptations and applications of the SLP model abound. For instance, Lee et al. (2011) modify the calculation of an air quality indicator by controlling for the preferential sampling process. In brief, air quality indicators are calculated as aggregates of observed values of various types of pollutions, measured over several observation sites, many of which may be assigned to regions with high pollution levels. Lee et al. fit the SLP model to various pollutant measurements to adjust estimated pollution surfaces for preferential sampling, thus calculating the indicator from the corrected surfaces. Notably, the authors find a substantial effect of preferential sampling on the air quality indicator.

A study concerning the impact of preferential sampling on health effect inference comes from Lee et al. (2015), again in the area of air pollution monitoring. The essence of the authors' approach is to model the placement of locations, treated as random quantities, then exposure given location, and finally health outcome given exposure. Briefly, the preferential sampling model used in this study is an adaptation of the SLP model, with a covariate driven mean of the response rather than a constant mean, among other structural changes to better match the context of air pollution monitoring. By simulation, the authors probed the exposure model to determine under what conditions preferential sampling has a substantial impact on the quality of the health effect estimates.

Gelfand et al. (2012) proposed a simulation based approach to delineate the effect of preferential sampling on the quality of predictive surfaces. The authors generated data under both a preferential sampling mechanism as well as a strategy in which observation sites were selected under complete spatial randomness (CSR). The authors found that the most accurate predictive surfaces were obtained when sampling was conducted under CSR, and that predicted surfaces were biased when the geostatisti-

cal model neglected to include sampling covariates, that is, covariates that influence the placement of observation sites.

The methods described hitherto have all primarily been developed with respect to the underlying application of environmental pollution monitoring, but are not, however, tailored toward most preferentially sampled disease surveillance datasets, which consist of spatially referenced case and control counts from diverse types of observation sites (e.g., farms, animal traps, etc), rather than a single smooth, normally distributed continuous response such as air or soil pollutants. Surprisingly, relatively few studies have developed methods to account for preferential sampling in disease surveillance, despite the potential for bias induced by this sampling mechanism. The few noted exceptions arise in veterinary health monitoring (Rinaldi et al., 2015; Cecconi et al., 2016).

A characteristic and illustrative example is found in Cecconi et al. (2016), who develop geostatistical model to predict the risk of parasitical infection in livestock. The authors propose a two-step Bayesian hierarchical model, wherein the adjustment for preferential sampling relies on incorporating estimated spatial sampling probabilities, obtained from the first step, as covariates into the linear predictor of their geostatistical model for disease risk in the second step. Formally, if the study region is partitioned into a grid of equally sized, non-overlapping cells, then the spatially varying sampling probability is estimated in a Binomial model:

$$K_j \sim \text{Binomial}(p_j, n_j)$$
$$\text{logit}(p_j) = \kappa + v_j + \epsilon_j$$

where K_j are the number of farms sampled in the jth grid cell, n_j are the total number of farms in the jth cell, and p_j is the sampling probability for the jth cell. In addition, p_j is logit-linear with intercept κ, along with a spatially structured residual v_j and spatially unstructured random error term ϵ_j. The authors choose an improper conditionally autoregressive (ICAR) prior distribution for v_j. To adjust for preferential sampling, the posterior estimates of the sampling probabilities are included as covariates in the linear component of the disease risk model, which is given by

$$
\begin{aligned}
Y_j &\sim \text{Bernoulli}(\pi_j) \\
\text{logit}(\pi_j) &= \gamma + s_j + x_j^T \beta + \alpha_j \tilde{p}_j
\end{aligned}
$$

where Y_j is an indicator for whether the jth sampled farm has livestock which test positive for the parasite, γ is an intercept, s_j are spatially structured residual terms arising from a mean zero, stationary Gaussian process, x_j are regression covariates, and $\tilde{p}_j$ are the posterior mean estimates of the sampling probabilities obtained from the first step. The use of $\tilde{p}_j$ as covariates here serve to adjust for preferential sampling, and were found to yield considerably different predictions of disease risk, as evidenced by a high observed Kullback-Leibler divergence calculated between risk estimates from the above model and one without inclusion of the sampling probabilities.

While this adjustment for preferential sampling has proven successful for veterinary health surveillance datasets, these applications hinge upon the ability to model the spatially varying sampling probabilities, p_j in the above notation, which are estimated from K_j, the number of sampled farms in a grid cell, and n_j, the total number of farms in the grid cell. However our methodological developments in this project focus on a scenario more common in zoonotic disease surveillance, wherein observation sites

aren't chosen from some fixed set of pre-existing locations (i.e. farms), but rather from animal specimen recovered via traps or opportunistically (e.g., roadkill). The quantities n_j and K_j from Cecconi et al. no longer have bearing under this mechanism of data collection. From a more abstract perspective, our proposed method does bear similarity to the approach by Cecconi et al. insofar as we include information from the distribution of observation sites into the part of our model describing case and control abundances. However, the way in which we incorporate this information more closely resembles the shared latent process model of Diggle et al. (2010).

Aside from the applications of air pollution monitoring and veterinary health, preferential sampling in species distribution modeling has gained recent attention (Conn et al., 2017; Gelfand and Shirota, 2019; Pennino et al., 2019). For instance, to address preferential sampling in this new domain, Pennino et al. (2019) propose a species distribution model resting upon the marked point process framework which originated in Diggle et al. (2010). However, to estimate this model the authors utilize the more computationally efficient approach of Rue et al. (2010) which, rather than relying on the Monte Carlo likelihood approximation of Diggle et al. (2010), involves inference by the approach of integrated nested Laplace approximation. A further point of divergence lies in the fact that Pennino et al. modeled the response of interest, i.e. observed species counts, as Gamma random variables, unlike the normally distributed response considered by Diggle et al. Highlighting the bias-inducing impact of preferential sampling, the authors show that the abundance of fish stock, specifically blue and red shrimp, is considerably overestimated in certain areas when preferential sampling goes unaddressed.

Further extending solutions for preferential sampling in the area of species distribution modeling, Gelfand and Shirota (2019) address a data fusion problem common in this domain, namely that of combining presence-only and presence-absence data.

Presence-only data refer to a data type in which only the locations, or presences, of a species are recorded, where presence-absence data explicitly detail both where a species was and was not observed. To jointly model presence-only and presence-absence data the authors propose an extension of the shared latent process framework of Diggle et al. (2010) and Pati et al. (2011). As in the shared latent process framework, the authors describe the distribution of presence-absence and presence-only sample sites with log Gaussian Cox processes:

$$\lambda_{PO}(s) = w^T(s)\beta_{PO} + \eta_{PO}(s)$$
$$\lambda_{PA}(s) = w^T(s)\beta_{PA} + \eta_{PA}(s)$$

where $\lambda_{PO}(s)$ is the intensity function of the point process describing the distribution of presence-only sample sites, and $\lambda_{PA}(s)$ is that of the presence-absence sites, while $\eta_{PO}(s)$ and $\eta_{PA}(s)$ are stationary spatial Gaussian processes specific to each data type. To correct for preferential sampling, η_{PO} and η_{PA} are included in the component of the model describing species presences, $Y(s) = I(Z(s) > 0)$, as

$$Z(s) = x^T(s)\alpha + \delta_{PA}\eta_{PA}(s) + \delta_{PO}\eta_{PO}(s) + w(s) + \epsilon(s)$$

with fixed effects $x^T(s)$, Gaussian process $w(s)$, and independent errors $\epsilon(s)$. Fitting their models to data detailing the distribution of invasive species in New England, the authors find that predictive performance suffers considerably when preferential sampling is ignored.

Lastly, we review a number of additional works addressing preferential sampling which

neither resemble the shared latent process model nor the disease surveillence model of Cecconi et al. Manceur and Kuhn propose a species occupancy model controlling for preferential sampling, based on the method of Bayesian image restoration, but one that relies on reliable knowledge of a control (reference) species. Preferential sampling was found under simulation to have a large impact on predicted species distributions. Chakraborty et al. (2010) build an areal level Bayesian hierarchical model for multi-species abundance, adjusting the abundance pattern to account for land use degradation and measurement error. Their areal level abundance data are in fact ordinal categorical variables, collected for multiple species. The authors constructed their degradation model on the latent scale, wherein an areal level spatial regression model induced correlation between abundances according to space (CAR model) and environmental covariates. But of the existing preferential sampling methods, Project 1 draws most heavily from the shared latent process model of Diggle et al. (2010) and its extension by Pati et al. (2011). We now introduce our proposed method, and show how it modifies these previous contributions to address preferential sampling in the context disease surveillance.

Proposed Method

Much existing research has focused on how preferential sampling impacts the prediction of spatial surfaces in applications such as air pollution monitoring, or increasingly, species distribution monitoring. However, a dearth of methodological development remains in the context of disease surveillance, with the exception of that pertaining to veterinary health (Cecconi et al., 2016), which, as we have seen, relies on certain structural assumptions arising from a fixed total number of sampling locations (i.e. farms). Yet in many real world surveillance systems, the data are in fact preferentially sampled, given that surveillance systems are often constrained by limited resources

to monitor for the disease primarily in high risk or high impact locations. The plague surveillance system undertaken by the California Department of Public Health, examined in the analysis chapter of this dissertation, is one such example. Given the impact of preferential sampling on spatial prediction shown in other applications, there is strong reason to suspect that similar issues may arise from preferentially sampled disease surveillance data.

The objective of our proposed model is to provide a better, less biased disease risk map obtained from preferentially sampled disease surveillance data. From a high level perspective, similar to the shared latent process model, our method has both locational and disease related components, where the locational component describes the pattern of observation in terms of a latent spatial process, which is shared with the component of the model describing the abundances of cases and controls. This sharing of the latent process serves to control for the effect of preferential sampling.

As in the shared latent process model of Diggle et al. (2010), we wish to describe the distribution of observation sites in terms of a spatial process. However, instead of directly working with a log Gaussian Cox process, we model the observation process in terms of a spatial regression problem. To that end, we begin by discretizing the study region into K non-overlapping, equally sized grid cells. Let indicator variables $\kappa_i \in \{0, 1\}$, $(i = 1, \ldots, K)$ denote whether the ith grid cell is observed by the disease surveillance system. We assume that κ_i, conditional on $\xi(x_i)$, is distributed as a Bernoulli random variable with spatially varying probability of success $\xi(x_i)$. Here we adopt the convention of referring to any spatial point in the study region as $x \in \mathbb{R}^2$, and of denoting the centroid of the ith grid cell as x_i. Therefore, by $\xi(x_i)$, we denote the value of the spatially varying probability of success, ξ, at the center point of the ith grid cell. Then, we model $\text{logit}(\xi(x_i)) = w(x_i)$, where $w(x_i)$ is the value of a mean-zero, stationary Gaussian process at the center point of the ith grid cell. We write

this Gaussian process as $w(x) \sim \mathcal{GP}(0, k(.,.;\theta,\phi))$ where k is a stationary, isotropic covariance function with spatial range θ and marginal variance ϕ. For the analysis and simulations conducted in this project we specify k to be an exponential covariance function. The portion of the model given until this point can be conceptualized as the locational component, describing the pattern of observation sites in the surveillance system.

Case and control counts of observed cells are denoted by Y_{im}, with $i \in \{h : \kappa_h = 1\}$ indexing grid cell identity and $m \in \{+,-\}$ distinguishing disease positive $(+)$ and negative $(-)$ status. Conditional on the value of the Gaussian process at point x_i, the Y_{im} are modeled as Poisson random variables with rates $\lambda_m(x_i)$. The rate function $\lambda_m(x)$ is log-linear in covariates $z_\lambda(x)^T\beta_m$ as well as $\alpha_m \times w(x)$, where $\alpha_m, m \in \{+,-\}$ is a scalar and $w(x)$ the value of the Gaussian process at point x. Here, for greater flexibility, we have allowed the covariate relationships β_m to vary by disease status. This model is written as

$$
\begin{aligned}
&\kappa_i|\xi(x_i) \sim \text{Bernoulli}(\xi(x_i)) \\
&\text{logit}(\xi(x)) = w(x) \\
&w(x) \sim \mathcal{GP}(0, k(.,.;\theta,\phi)) \\
&Y_{im}|w(x_i) \sim \text{Poisson}(\lambda_m(x_i)) \\
&\log(\lambda_m(x)) = z_\lambda(x)^T\beta_m + \alpha_m \times w(x)
\end{aligned}
\tag{1.1}
$$

The inclusion of $w(x)$ in the mean function of Y_{im} induces stochastic dependence between disease case or control counts and survey location. Similar to the model proposed by Pati et al. (2011), the shared latent process $w(x)$ is multiplied by parameters

$\alpha_m, m \in \{0, 1\}$, intended to govern the strength of preferential sampling. That is, the greater the magnitude of α_m, the greater the tendency of Y_{im} to take higher values at observed locations than unobserved ones. However, unlike Pati et al. we here specify not one but two α parameters, α_+ and α_-, corresponding to the effects of preferential sampling on the observed abundances of cases and controls, respectively. These two parameters are crucial for capturing the tendency of the observation process to focus on areas of elevated disease risk. To see this, recall that the disease odds at point x are given by the ratio of case and control intensity functions, $\lambda_+(x)/\lambda_-(x)$, and thus, disease log odds at point x are equal to

$$z_\lambda(x)^T\beta_+ + \alpha_+ \times w(x) - z_\lambda(x)^T\beta_- - \alpha_- \times w(x)$$

Hence, for fixed values of β_+, β_-, and $w(x)$, disease log odds, and thus, disease risk, increase as $\alpha_+ \times w(x) - \alpha_- \times w(x)$ increases. A further noteworthy distinction of model (1.1) from the existing preferential sampling methods is the fact that the responses measured, Y_{im}, are Poisson, rather than normally, distributed. Consequently, the reliance on Gibbs sampling to update $w(x)$ is no longer feasible here. The next section details the strategy used to fit model (1.1).

1.2.6 Model Fitting

Model (1.1) is fit by a Markov Chain Monte Carlo solution consisting of separate Hamiltonian Monte Carlo samplers for $\alpha_+, \alpha_-, \beta_+, \beta_-$ and spatial random effects w, along with a Metropolis-Hastings random walk update for the spatial range θ, and finally with a Gibbs sampling update for the marginal variance ϕ. The samplers involved have been implemented from scratch in the R statistical programming lan-

guage, version 3.4.3, without making use of pre-built MCMC packages due to the unique nature of the model being fit.

We assign normal priors to $\alpha_+, \alpha_-, \beta_+$ and β_-. In the analyses and simulations conducted here we assign uninformative priors for β_+ and β_-, with large prior variances, as is often the case when estimating slope parameters in Bayesian analysis. The spatial range parameter θ of the exponential covariance function was estimated via Metropolis-Hastings random walk. We note that, due to the constraint of $\theta > 0$, the proposal distribution used to generate a proposed next value for θ was specified as the log-normal distribution, which has density function

$$q(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(- \frac{(\log(x) - \mu)^2}{2\sigma^2} \Big)$$

for $x > 0$. The mean of the proposal distribution was taken to be the log of the current value of θ. That is, given the gth MCMC sample $\theta^{(g)}$, the proposed next value in the Markov chain was distributed as

$$\theta^{(g+1)} \sim \text{Log-Normal}(\log(\theta^{(g)}), \sigma^2)$$

The proposal standard deviation σ^2 was manually tuned to yield acceptance rates close to 0.5. The acceptance probability was calculated as

$$\min\Big(1, \frac{\ell(w^{(g)}; \theta^{(g+1)})}{\ell(w^{(g)}; \theta^{(g)})} \times \frac{p(\theta^{(g+1)})}{p(\theta^{(g)})} \times \frac{q(\theta^{(g)})}{q(\theta^{(g+1)})}\Big)$$

where $\ell(w^{(g)}; \dots)$ is the log likelihood of the spatial random effects w at the gth MCMC iteration, and $p(\dots)$ is the prior density of θ, taken here to be the gamma

distribution, and $\frac{q(\theta^{(g)})}{q(\theta^{(g+1)})}$ is the ratio of log-normal densities from the current and proposed values of θ, which would have cancelled out had the proposal distribution been symmetric.

We assign an Inverse-Gamma prior distribution to the spatial marginal variance ϕ, in order to make use of the fact that the conditional distribution of ϕ given the random effects w is also Inverse-Gamma. Specifically,

$$\phi|w \sim 1/\text{Gamma}(N/2 + a, w^T H^{-1} w + b)$$

where N is the number of elements of w, a and b are the shape and scale parameters of the prior distribution of ϕ, and H is the correlation matrix of the random effects w. Consequently, ϕ can be updated at each step of the Markov chain by drawing a sample from $\phi|w$, rather than relying on a more computationally expensive Metropolis-Hastings or Hamiltonian Monte Carlo sampler.

In contrast, due to the fact that, in the realm of disease surveillance, the measured responses Y_{im} are Poisson distributed counts, rather than normally distributed real numbers such as air pollution measurements, the spatial random effects w have no analytic form for their posterior conditional distribution given Y. Consequently, the random effect vector w is updated via a Hamiltonian Monte Carlo (HMC) sampler. This technique was chosen given its effectiveness for updating high dimensional, spatially structured parameters. Briefly, HMC proposes new parameter states by simulating the dynamics of Hamiltonian physics, which describes the total energy of a system as the sum of potential and kinetic energies. Here, the potential energy of the system is taken as the negative log likelihood of the current parameter state, and a new parameter state is reached by evaluating the gradient of the potential energy. Thus, by incorporating information from the gradient of the log likelihood in

its proposal step, HMC is able to explore the parameter space more efficiently and, crucially, account for spatial correlation between elements of the parameter vector, a distinction which would not hold if a Metropolis-Hastings random walk updating strategy were employed here. Hamiltonian Monte Carlo algorithms are parametrized by a step size parameter, related to the degree of change undertaken in the proposal step, and length parameter, which controls the number of iterations for which Hamiltonian dynamics are simulated in each proposal step. The step size parameter was automatically tuned by the strategy of dual averaging, presented by Hoffman and Gelman (2010), which alters the step size after each proposal step based on a convex optimization algorithm which compares the current acceptance probability with the desired acceptance rate. The length parameter was manually tuned.

Hamiltonian Monte Carlo samplers were also assigned to separately update $\alpha_+, \alpha_-, \beta_+,$ and β_-. Despite the fact that these parameters are low dimensional and spatially uncorrelated, HMC sampling showed less inter-sample correlation than Metropolis-Hastings random walk and so was ultimately preferred. A more detailed technical description of HMC along with dual averaging is presented in the next section.

1.2.7 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC), introduced by Duane et al. (1987), is an improvement over conventional MCMC samplers insofar as its proposals are influenced by the gradient of the log density of the parameter of interest. Consequently, by avoiding random walk behavior, HMC more efficiently explores the parameter space, and is also less sensitive to correlated parameters. For these reasons HMC is particularly adept at updating high dimensional parameters, which is of key importance in fitting spatial hierarchical models.

The following technical description of the traditional Hamiltonian Monte Carlo algorithm is based on the development by Hoffman and Gelman (2014), who also provide a strategy to adaptively-tune the Hamiltonian Monte Carlo sampler, which we examine later. First, a momentum variable r_d is introduced for each model parameter θ_d, solely for the purpose of simulating Hamiltonian dynamics, and is typically drawn from the standard normal distribution. Let $\mathcal{L}(\theta)$ be the log density of the parameters of interest θ. Then the joint density of θ and momentum r is, up to a constant, given by

$$p(\theta, r) = \exp(\mathcal{L}(\theta) - 0.5r \cdot r)$$

This model can be interpreted as a Hamiltonian system where θ is the position of a particle, $\mathcal{L}$ is the negative potential energy of the particle, $0.5r \cdot r$ is the kinetic energy of the particle, and $\log p(\theta, r)$ is the particle's negative energy. Each iteration of the sampler describes the change in the position of the particle over time according to Hamiltonian dynamics. Specifically, updates are carried out according to the Störmer-Verlet "leapfrog" integrator:

$$\begin{aligned} r^{t+\epsilon/2} &= r^t + (\epsilon/2)\nabla_\theta \mathcal{L}(\theta^t) \\ \theta^{t+\epsilon} &= \theta^t + \epsilon r^{t+\epsilon/2} \\ r^{t+\epsilon} &= r^{t+\epsilon/2} + (\epsilon/2)\nabla_\theta \mathcal{L}(\theta^{t+\epsilon}) \end{aligned}$$

where t indexes the time of the position and momentum of the particle, ∇_θ is the gradient with respect to θ, and ϵ is the step size specified a priori. For each Hamiltonian Monte Carlo sample drawn, first the momentum r variable is sampled from

a standard multivariate normal distribution. Then, L leapfrog updates are applied to the position and momentum variables to put forward proposal values for θ and r, denoted $\tilde{\theta}$ and $\tilde{r}$. Lastly, $\tilde{\theta}$ and $\tilde{r}$ are accepted or rejected by a Metropolis step with acceptance probability given by $p(\tilde{\theta}, \tilde{r})/p(\theta, r)$. This Hamiltonian Monte Carlo algorithm was custom implemented from scratch in R for this dissertation.

While HMC more efficiently explores the parameter space and is less sensitive to correlated parameters than the Metropolis-Hastings random walk sampler, its more widespread adoption has been encumbered by the difficulty of setting the simulation length parameter L and step size ϵ. Setting ϵ too high may result in an overly low acceptance rate, while setting it too low produces the opposite effect. Setting L too low may result in random walk behavior by generating samples which are too close together. Likewise, high values of L may achieve the same effect by generating particle paths which loop back toward their initial states.

Several adaptive MCMC tuning techniques (e.g., Andrieu and Thoms, 2008) lend themselves to setting the step size parameter ϵ. To fit model (1.1) by Hamiltonian Monte Carlo, this dissertation makes use of the dual averaging method proposed by Hoffman and Gelman (2014), wherein ϵ is iteratively updated to achieve a target acceptance rate, recommended to be approximately 0.65. Hoffman and Gelman adapt the dual averaging scheme proposed by Nesterov (2009), originally as a means for convex optimization, by letting $H_t = \alpha_t - \delta$, where α_t is the acceptance probability of the tth HMC sample and δ is the desired acceptance rate. The problem then is to update ϵ such that $E_t[H_t|\epsilon] = 0$. To achieve this end, the following updates are applied:

$$\epsilon_{t+1} \leftarrow \mu - \frac{\sqrt{t}}{\gamma}\frac{1}{t+t_0}\Sigma_{i=1}^{t} H_i$$

$$\bar{\epsilon}_{t+1} \leftarrow \eta_t \epsilon_{t+1} + (1-\eta_t)\bar{\epsilon}_t$$

where μ is a point to which values of ϵ_t are shrunk, $\gamma > 0$ controls the amount of shrinkage toward μ, and t_0 is a stabilizing parameter. For fitting the models in this dissertation values of $\mu = \log(10 * 0.05), \gamma = 0.05$, and $t_0 = 10$ were chosen, as suggested by Hoffman and Gelman (2014). The success of the dual averaging technique is ensured by the fact that the sequence of $\bar{\epsilon}_{t+1}$ values is guaranteed to converge to a value such that $E_t[H_t|\epsilon] = 0$ converges to zero. Thus, in practice, ϵ values are updated at each iteration of the HMC sampler for a tuning period to allow for this convergence, typically for a predetermined tuning period, after which the step size ϵ is fixed at the final iterate of $\bar{\epsilon}_t$.

Hoffman and Gelman also propose a method to avoid manually setting the length parameter L, known as the NUTS (No U-turn) sampler. The basic idea is to simulate Hamiltonian dynamics each iteration for as many steps until the trajectory of proposed value starts to turn back toward its initial state. However, for the majority of datasets fit in this dissertation, model (1.1) was fit with values of L set between 8 and 10, due to the fact that the added runtime associated with adaptively tuning L proved to outweigh the cost of manual tuning, especially at high resolution. Thus, L was manually set while ϵ was adaptively tuned for each HMC sampler associated with model (1.1).

1.2.8 Spatial Downscaling

Up until this point we have defined our proposed method to adjust for preferential sampling in estimating disease risk maps, and discussed the model fitting process using a combination of MCMC techniques, i.e. Hamiltonian Monte Carlo, Metropolis-Hastings random walk, and Gibbs sampling. However, one very salient consideration has been omitted hitherto, and that is the problem of efficiently fitting the model at high resolution. Recall that model (1.1) contains a Gaussian process to explain the distribution of observation sites as well as additional variation in the observed abundances of cases and controls. The study region is discretized, so that the pattern of observation sites is encoded by way of indicator variables representing whether each discretization cell is observed. Consequently, the Gaussian process realizes as many random effects as there are grid cells in the study region. For large study regions at fine levels of discretization, the size of the covariance matrix of the Gaussian process can become so great as to prohibit its inversion. For instance, a disease surveillance system wishing to estimate a risk map over the state of California at a 4 km^2 resolution entails a covariance matrix of 25,701 rows, which cannot be efficiently inverted quickly enough to make fitting model (1.1) feasible in a reasonable period of time.

Fortunately, a rich body of solutions exists for fitting spatial models with Gaussian processes at low resolutions (e.g. Wikle and Cressie, 1999), and then subsequently extending the model to higher resolutions, an approach referred to as *spatial downscaling*. In many approaches, the Gaussian process may be approximated by realizations of random effects taken at a series of "knots", or points covering the study region, which are typically evenly spaced and far enough apart so as to make computation feasible. The value of the Gaussian process at points between the knots is then interpolated by a variety of means. One straightforward approach is to simply krige the values of the Gaussian process to unobserved points at lower resolution, and in spe-

cial cases, use a tapered covariance matrix to ease the inversion process (Furrer et al., 2006). The method of tapering assigns zeros to the covariance matrix for distances above a certain value, in order to induce sparsity. However, in the analysis portion of this dissertation, tapering was found to be an inadequate solution given that the tapered covariance matrix still had over 440,000,000 nonzero elements. Other solutions for downscaling the Gaussian process abound. Reich et al. (2014) develop a spatial downscaler making use of the spectral representation of Gaussian processes, while Nychka et al. (2015) develop the method of lattice kriking, using nested tapered bivariate splines to predict values of a Gaussian process at multiple levels of resolution.

Our approach is to fit model (1.1) at a coarser resolution, assume the covariate relationships $z_\lambda(s)^T\beta_m$ along with $\alpha_m \times w(s)$ hold at higher resolutions, and subsequently use the known high resolution covariate values of $z_\lambda(s)$ along with values of the spatial field $w(s)$ downscaled via spline interpolation to predict case and control abundances at higher resolution. Specifically, let $s_1, \ldots, s_n$ be the center points of the cells which form the low resolution discretization of the study region, and suppose $\hat{\beta}_m, \hat{\alpha}_m, m \in (+,-), \hat{w}(s_1), \ldots, \hat{w}(s_n), \hat{\theta}$ and $\hat{\phi}$ are the estimates obtained from fitting model (1.1) at low resolution. Then, if the high resolution discretization of the study region contains cell center points $s'_1, \ldots, s'_h$, the estimated intensities for the ith center point are $\log(\hat{\lambda}_m(s'_i)) = z_\lambda(s'_i)^T\hat{\beta}_m + \hat{\alpha}_m \times \tilde{w}(s'_i)$, where $\tilde{w}$ is the estimated value of the Gaussian process obtained by interpolating $\hat{w}(s_1), \ldots, \hat{w}(s_n)$ via thin plate splines. Spline interpolation was carried out by the *tps* function of the fields package in *R*. The benefit of such an approach is that is makes use of the known high resolution covariate values in the PRISM dataset, which would have been ignored if a cruder approach were adopted, such as simply smoothing the estimated risk surface fit at low resolution. We illustrate the results of our process by the downscaled risk map of plague in Sciurids (the rodent family of squirrels) in the state of California (Figure

1.3), with the understanding that we will review these estimates in more detail in the analysis section below.

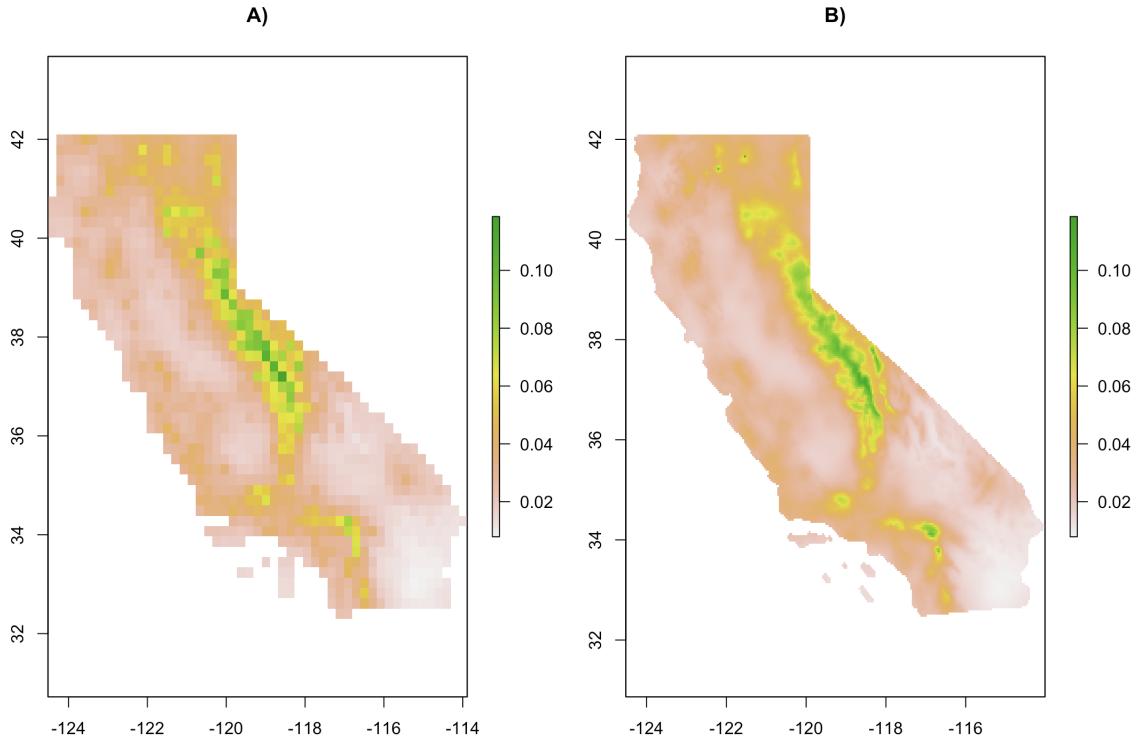

Figure 1.3: Low and high resolution plague risk map comparison. A) 426 km^2 plague risk map B) Spatially downscaled 16 km^2 plague risk map.

1.3 Simulation 1: Comparative Performance

1.3.1 Introduction

This simulation study assesses performance of the proposed preferential sampling model in comparison to existing, benchmark methods which do not account for the sampling process. To form a more complete view of comparative performance under different real world scenarios, we evaluate our models over simulated datasets en-

compassing a range of strengths of preferential sampling. The intent is to determine the circumstances, if any, under which the proposed method outperforms the benchmarks, as well as to quantify the magnitudes of any gains in performance offered by the method we have introduced.

Models Compared

In this study our proposed preferential sampling model is compared against benchmark methods which do not attempt to account for the sampling process that gave rise to the data. The reference methods we consider consist of a spatial Poisson model (1.2) and non-spatial Poisson regression model (1.3).

As described in the methods section, we discretize the study region into $l = 1, \ldots, L$ equally sized, nonoverlapping grid cells, and suppose that $i = 1, \ldots, h$ of these cells have been observed by the surveillance system. The spatial Poisson model (1.2) assumes case and control counts of observed grid cells, denoted Y_{i+} and Y_{i-}, representing the count of cases or controls in the ith grid cell, respectively, follow independent Poisson distributions $Y_{im} \sim \text{Poisson}(\lambda_m(x_i))$ for $m \in \{+, -\}$ and $x_i \in \mathbb{R}^2$ taken to be the coordinates of the centroid of the ith observed grid cell. The rate functions $\lambda_m(x)$ are log-linear in $z_\lambda(x)^T\beta_m + w_m(x)$. Here, $z_\lambda(x)$ are disease related spatial covariates and β_m are parameters distinguished by disease status $m \in \{+, -\}$. The intensity functions of cases ($m = +$) and controls ($m = -$) also incorporate independent mean zero Gaussian processes $w_m(x)$, whose covariance functions are assumed to be exponential and parametrized by range θ_m and marginal variance ϕ_m, allowing the range and marginal variance to differ by disease status m. The inclusion of these Gaussian processes is intended to model spatial variation in the responses which cannot be explained by the deterministic trends $z_\lambda(x)^T\beta_m$. This spatial Poisson model is summarized as

$$
\begin{aligned}
Y_{im} &\sim \text{Poisson}(\lambda_m(x_i)) \\
\lambda_m(x) &= \exp(z_\lambda(x)^T \beta_m + w_m(x)) \\
w_m(x) &\sim \mathcal{GP}(0, k(.,.;\theta_m, \phi_m))
\end{aligned}
\tag{1.2}
$$

Model (1.2) is fit by MCMC using Hamiltonian Monte Carlo samplers for w_m and β_m, Gibbs samplers for ϕ_m, and Metropolis-Hastings random walks for θ_m. Having obtained estimates for $\beta_m, \theta_m, \phi_m$, and w_m at observed sites, it remains to predict the values of the latent processes w_m at unobserved sites, in order to calculate disease risk over the entire study area. This prediction is formed by Bayesian kriging, the results of which, along with previously obtained parameter estimates for β_m, allow estimation of case and control intensities, $\hat{\lambda}_m$, anywhere in the study region. Estimated disease odds are then obtained as the ratio of intensities, $\hat{\lambda}_+/\hat{\lambda}_-$.

The next reference method we consider entails non-spatial Poisson regression models which regress case and control counts on spatial covariates alone (1.3). Here case ($m = +$) or control ($m = -$) counts of the ith observed grid cell follow Poisson distributions with rate $\lambda_m(x_i)$, where x_i denotes the center point of the ith cell. Rates λ_m are log-linear in disease covariates $z_\lambda(x)^T \beta_m$, where β_m are parameters specific to case or control status. Unlike model (1.2), no spatial processes are included in the intensity functions, allowing estimation of the disease odds λ_+/λ_- to be conducted immediately after estimates for β_m are obtained. This model is summarized as

$$Y_{im} \sim \text{Poisson}(\lambda_m(x_i)) \tag{1.3}$$
$$\lambda_m(x) = \exp(z_\lambda(x)^T \beta_m)$$

Evaluation Metrics

The primary evaluation metric of models (1.1), (1.2), and (1.3) is root mean squared error in estimated log disease odds. True disease log odds of any cell x are calculated as $o_i = \log(\lambda_+(x)/\lambda_-(x))$, where $\lambda_+(x)$ and $\lambda_-(x)$ are case and control intensities, respectively, evaluated at the center point of cell x. RMSE is then calculated as

$$K^{-1} \sum_{i=1}^{K} \sqrt{(\hat{o}_i - o_i)^2}$$

for a study region with K grid cells and estimated log odds $\hat{o}_i = \log(\hat{\lambda}_+(x)/\hat{\lambda}_-(x))$.

1.3.2 Data

This simulation study evaluates the performance of models (1.1), (1.2), and (1.3) with respect to a total of 50 datasets, 25 of which were simulated under a low level of preferential sampling, and 25 of which were simulated under a high level. The level or strength of preferential sampling was quantified in terms of the average percentage of the linear predictor of log odds that was constituted by the sampling related terms of model (1.1), $\alpha_+ w(x) - \alpha_- w(x)$. Specifically, recall that log disease odds for the grid cell with centroid at point $x_i \in \mathbb{R}^2$ are given by the difference of case and control log intensities:

$$z_\lambda(x_i)^T\beta_+ + \alpha_+ w(x_i) - z_\lambda(x_i)^T\beta_- - \alpha_- w(x_i)$$

We conceptualize two contributions to log odds, one arising from fixed effects, $z_\lambda(x_i)^T\beta_+ - z_\lambda(x_i)^T\beta_-$, and another due to preferential sampling, $\alpha_+ w(x_i) - \alpha_- w(x_i)$. The percent contribution attributed to preferential sampling for the ith cell is then

$$p_i = \frac{|\alpha_+ w(x_i) - \alpha_- w(x_i)|}{|\alpha_+ w(x_i) - \alpha_- w(x_i)| + |z_\lambda(x_i)^T\beta_+ - z_\lambda(x_i)^T\beta_-|}$$

To to form an overall estimate of the strength of preferential sampling for a given simulated dataset, we average the percent contributions across all K grid cells of the study region, $\bar{p} = K^{-1}\Sigma_{i=1}^K p_i$. If $\bar{p}_d$ is the average percent contribution for the dth simulated dataset ($d = 1, \ldots, 25$), then to represent a low level of preferential sampling, simulation parameters β_m and α_m ($m \in \{+, -\}$) were chosen so that the median of the $\bar{p}_d$ ($d = 1, \ldots, 25$) fell at 13.22% (mean 14.68%, maximum 29.67%). Thus, in our construction, a low level of preferential sampling corresponds to a median preferential sampling contribution of 13.22% over 25 simulated datasets. Similarly, to represent a high level of preferential sampling, β_m and α_m ($m \in \{+, -\}$) were tuned so that the median preferential sampling contribution was 34.26% (mean 36.28%, maximum 59.38%). The specific parameterizations giving rise to these percentages were, for low preferential sampling, $\alpha_+ = 0.5, \alpha_- = 0.3, \beta_+ = (1.00, 0.75, 0.25)^T$, and $\beta_- = (3.0, 1.0, 0.5)^T$. High preferential sampling was simulated with $\alpha_+ = 1, \alpha_- = -0.5, \beta_+ = (-1.50, 0.25, 0.25)^T$, and $\beta_- = (3.5, 1.0, 0.5)^T$. The vector of spatial random effects, $w = (w_1, \ldots, w_K)$, was re-simulated for every dataset from a Gaussian process with mean zero and exponential covariance function parametrized by a range $\theta = 7$ and marginal variance $\phi = 12$. Fixed effect covariates $z_\lambda(x_i)^T$ were taken as the first and second principal components of the PRISM climatic dataset, along with

the inclusion of an intercept (Figure 1.4).

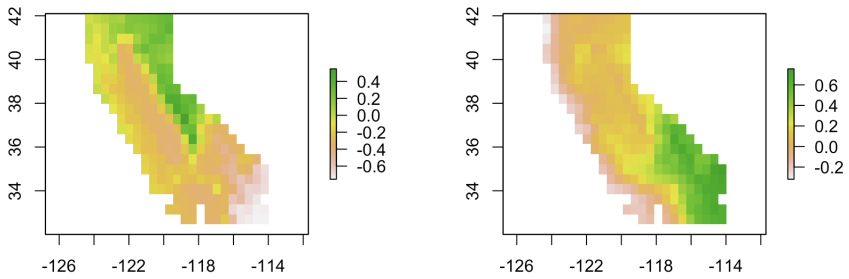

Figure 1.4: A) First and B) second covariate surfaces used to simulate disease surveillance datasets over a study region discretized into 259 grid cells.

Dataset generation begins with simulation of the spatial random effect vector $w = (w_1, \ldots, w_K)$ from a multivariate normal distribution. For each dataset the study region was discretized into $K = 259$ nonoverlapping grid cells, a granularity reached by a tradeoff between computational efficiency and breadth of spatial extent. Given the simulated vector of random effects w, indicators κ_i were then simulated from model (1.1), representing the distribution of observation sites generated by the disease surveillance process. Contingent upon this pattern of observation and the random effect vector w, case and control counts were then simulated from model (1.1). Thus, each simulated dataset consists of a different set of observation sites and case/control counts. Under a low level of preferential sampling there were on average 114.56 observed grid cells (standard deviation: 71.953) (Table 1.1), throughout which an average disease prevalence of 0.153 (standard deviation: 0.053) persisted. Datasets with high preferential sampling averaged 136.84 observed cells (standard deviation: 60.55) with an average disease prevalence of 0.255 (standard deviation: 0.293).

Sampling	Quantity	Mean	SD	Q1	Q2
Low	Observed Cells	114.56	71.953	47	178
Low	Prevalence	0.153	0.053	0.125	0.177
High	Observed Cells	136.84	60.55	95	183
High	Prevalence	0.255	0.293	0.014	0.362

Table 1.1: Summary of observed cells and disease prevalences for surveillance datasets simulated under different levels of preferential sampling.

1.3.3 Model Fitting

Model (1.1)

For each of the 50 datasets model (1.1) was fit by the MCMC implementation described in the methods section. The length tuning parameters for each Hamiltonian Monte Carlo sampler were set to 8, while sampler tuning periods were fixed at 2,000, with target acceptance rates of 0.65 as recommended by Hoffman and Gelman (2014). The proposal standard deviation of the MHRW sampler was set to 0.15, a value tending to yield acceptance rates close to 0.50. MCMC initial values were assigned using the calibrated initialization strategy described in the methods section. Given heuristic estimates $\hat{\alpha}_+, \hat{\alpha}_-, \hat{\theta}$, and $\hat{\phi}$ obtained from the initialization strategy, the following priors were specified:

$$\alpha_+ \sim N(\hat{\alpha}_+, 3)$$

$$\alpha_- \sim N(\hat{\alpha}_-, 3)$$

$$\theta \sim \text{Gamma}(\text{shape}(\hat{\theta}), \text{scale}(\hat{\theta}))$$

$$\phi \sim \text{Inverse-Gamma}(\text{shape}(\hat{\phi}), \text{scale}(\hat{\phi}))$$

where $\text{shape}(\hat{\theta})$ and $\text{scale}(\hat{\theta})$ were calculated so as to equate the prior mean to the

heuristic estimate $\hat{\theta}$ and prior variance to 2. Similarly shape($\hat{\phi}$) and scale($\hat{\phi}$) were calculated so as to result in a prior mean equal to the heuristic estimate $\hat{\phi}$ and prior variance of 2. Priors for β_+ and β_- were specified as $N(0, 100)$. Given these prior distributions and the aforementioned tuning parameters, model (1.1) was fit to each dataset with a total of 10,000 MCMC samples with a burnin of 3,000.

Model (1.2)

The spatial Poisson model was fit by a custom-built MCMC implementation wherein the spatial random effect vectors $w_+(x)$ and $w_-(x)$, along with β_+ and β_-, were updated via Hamiltonian Monte Carlo (HMC) samplers, while the spatial range θ was updated by Metropolis-Hastings random walk and marginal variance ϕ by Gibbs sampling. HMC step sizes were algorithmically tuned to yield target acceptance rates close to 0.65 under a tuning period of 2,000 samples, while the length parameters of the HMC samplers were fixed at 8, a value chosen based on manual evaluation. To update θ, the proposal standard deviation for the random walk was set to 0.3. The following prior distributions were assigned:

$$\beta_m \sim N(0, 100), m \in \{+, -\}$$
$$\theta_m \sim \text{Gamma}(\text{shape}(7), \text{scale}(7)), m \in \{+, -\}$$
$$\phi_m \sim \text{Inverse-Gamma}(3, 40), m \in \{+, -\}$$

Here uninformative priors have been specified for β_+, β_-, ϕ_+ and ϕ_-, while the range parameters θ_m have been assigned priors whose shape and scale parameters were chosen to yield prior means of 7, the true value of θ, and variances of 5. The model

was fit to each dataset with a total of 10,000 samples and burnin period of 3,000.

Model (1.3)

Model (1.3) was fit by maximum likelihood, as implemented in the *glm* package of the R statistical software version 3.4.3.

1.3.4 Results

Under a low level of preferential sampling the proposed model had the lowest average root mean squared error in estimated log disease odds, at a value of 0.573 (standard deviation: 1.449), followed by the Poisson model with an average RMSE of 0.706 (standard deviation: 0.436), and lastly by the spatial Poisson model (average RMSE: 0.728, standard deviation: 1.024) (Table 1.2). It is noteworthy that the RMSEs for the proposed model and model (1.2) inflated to considerably high values under certain datasets (Figure 1.5A). Further inspection reveals a strong relationship between RMSE in the proposed model and the number grid cells observed by the sampling process for each simulated dataset. In Figure (1.6A) we see that RMSE in the proposed model spikes up for datasets with roughly fewer than 100 observations, especially so for datasets simulated under a high degree of preferential sampling. Inspection of the relationship between RMSE and disease prevalence of the simulated dataset also shows an increase in RMSE as prevalence decreases (Figure 1.6B). If we restrict our evaluation to datasets with at least 75 observed grid cells we see that the proposed model has an average RMSE almost half that of the spatial Poisson model and less than a third of that of the Poisson model (Table 1.2).

As the strength of preferential sampling increased from a low to high level, the differ-

ences in RMSEs among the three models increased sharply. Under a high degree of preferential sampling, when restricted to simulated datasets with at least 75 observed grid cells, the proposed model still had the lowest average RMSE at 1.061, over 30% lower than the average RMSE of the spatial Poisson model (mean: 1.645), and considerably lower than the Poisson model (mean: 3.796). Figure (1.5) summarizes the differences in RMSE of the three models across different levels of preferential sampling, and across differing sample sizes. We see that as the strength of preferential sampling increases, and as we restrict our evaluation to datasets with a sufficient number of observed grid cells, the RMSEs in estimated log odds of the reference models inflate relative to that of the proposed method.

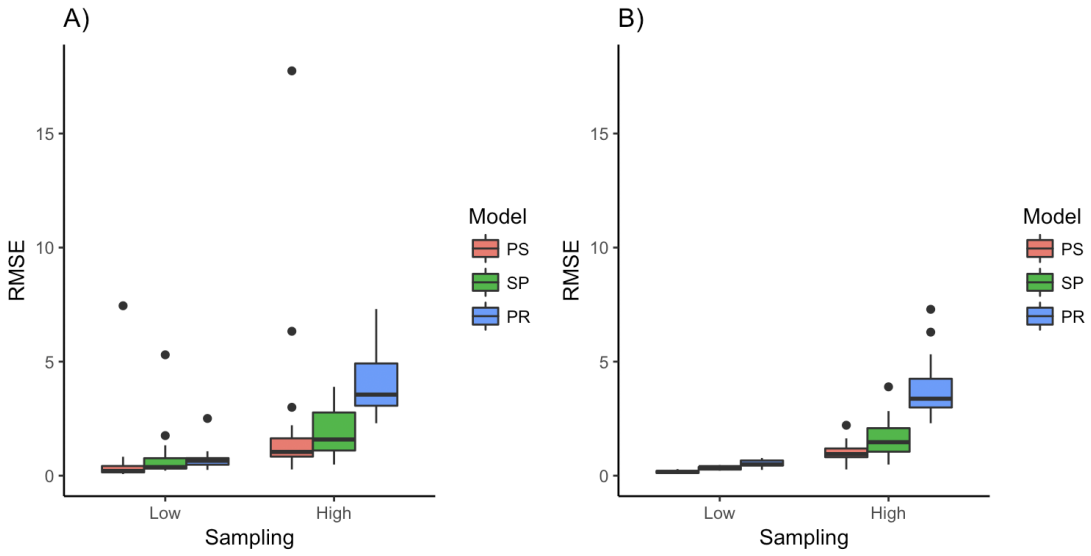

Figure 1.5: Root mean squared errors in estimated log disease odds under low and high levels of preferential sampling for the preferential sampling model (PS), spatial Poisson model (SP), and Poisson regression model (PR) when considering A) all simulated datasets and B) only simulated datasets with at least 75 observed grid cells.

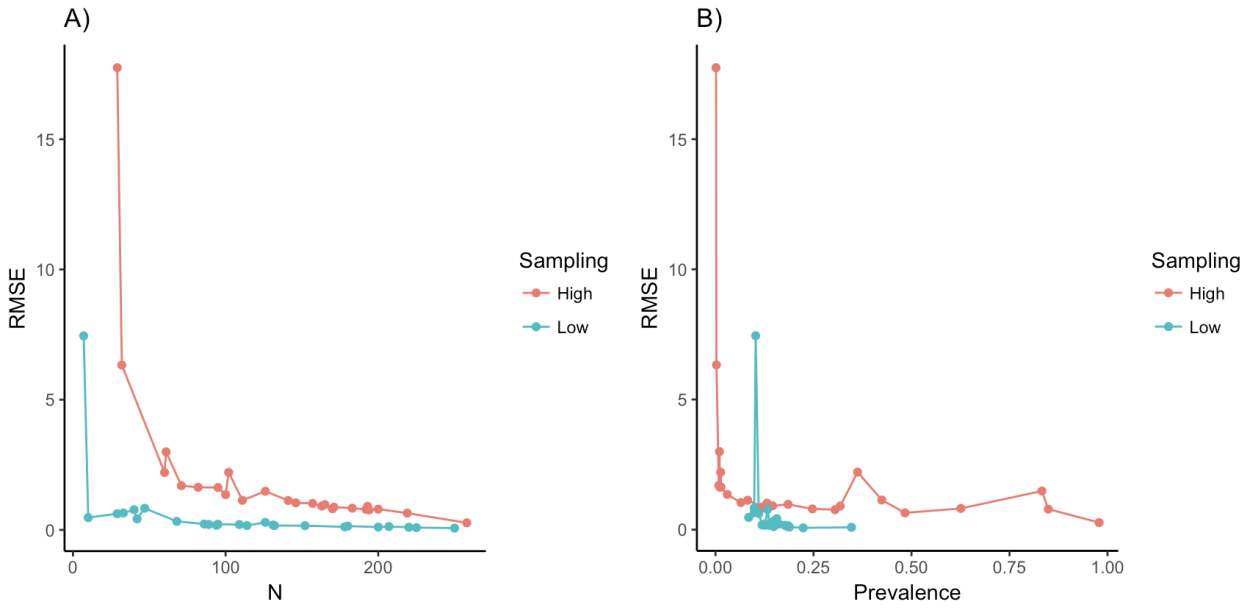

Figure 1.6: Root mean squared errors in estimated log disease of the proposed model versus A) the number of observed cells and B) the disease prevalence of the simulated dataset.

Sampling	Model	Avg RMSE	Sd RMSE	Avg RMSE ($N \geq 75$)	Sd RMSE ($N \geq 75$)
Low	PS	0.573	1.449	0.163	0.057
Low	SP	0.728	1.024	0.336	0.069
Low	PR	0.706	0.436	0.535	0.156
High	PS	2.088	3.469	1.061	0.428
High	SP	1.951	0.989	1.645	0.818
High	PR	4.08	1.417	3.796	1.323

Table 1.2: Summary of RMSE in estimated log disease odds under low and high levels of preferential sampling for the proposed model (PS), spatial Poisson model (SP), and poisson regression model (PR). Avg and Sd RMSE refer to the mean and standard deviation in RMSE, while ($N \geq 75$) denotes restricting the calculation of RMSE to datasets in which at least 75 grid cells were observed.

While the above comparisons showed that the proposed model tended to have lower RMSE in estimated log disease odds than the reference approaches, especially under a high level of preferential sampling, we still wish to take a more detailed view of the performance of our proposed model by examining biases in individual model parameters. We begin with α_+ and α_-, the parameters modulating the effect of preferential sampling on the abundances of cases and controls, respectively. When

considering all simulated datasets regardless of the number of observed grid cells, under low preferential sampling, average biases for α_+ and α_- were -0.087 and -0.024, respectively. When only simulated datasets with at least 75 observed grid cells were considered, the mean biases for α_+ and α_- under low preferential sampling decreased to 0.019 and 0.006. For a high level of preferential sampling, mean biases in α_+ and α_- were 0.239 and -0.029, but only 0.068 and -0.029 over simulated datasets with at least 75 observed grid cells. Biases in α_+ and α_- over differing levels of preferential sampling and sample size requirements are summarized in Figure (1.7A,B). For both levels of preferential sampling and both parameters α_+ and α_-, biases showed a general trend of rapid decrease as the number of observed grid cells increased (Figure 1.7C, D).

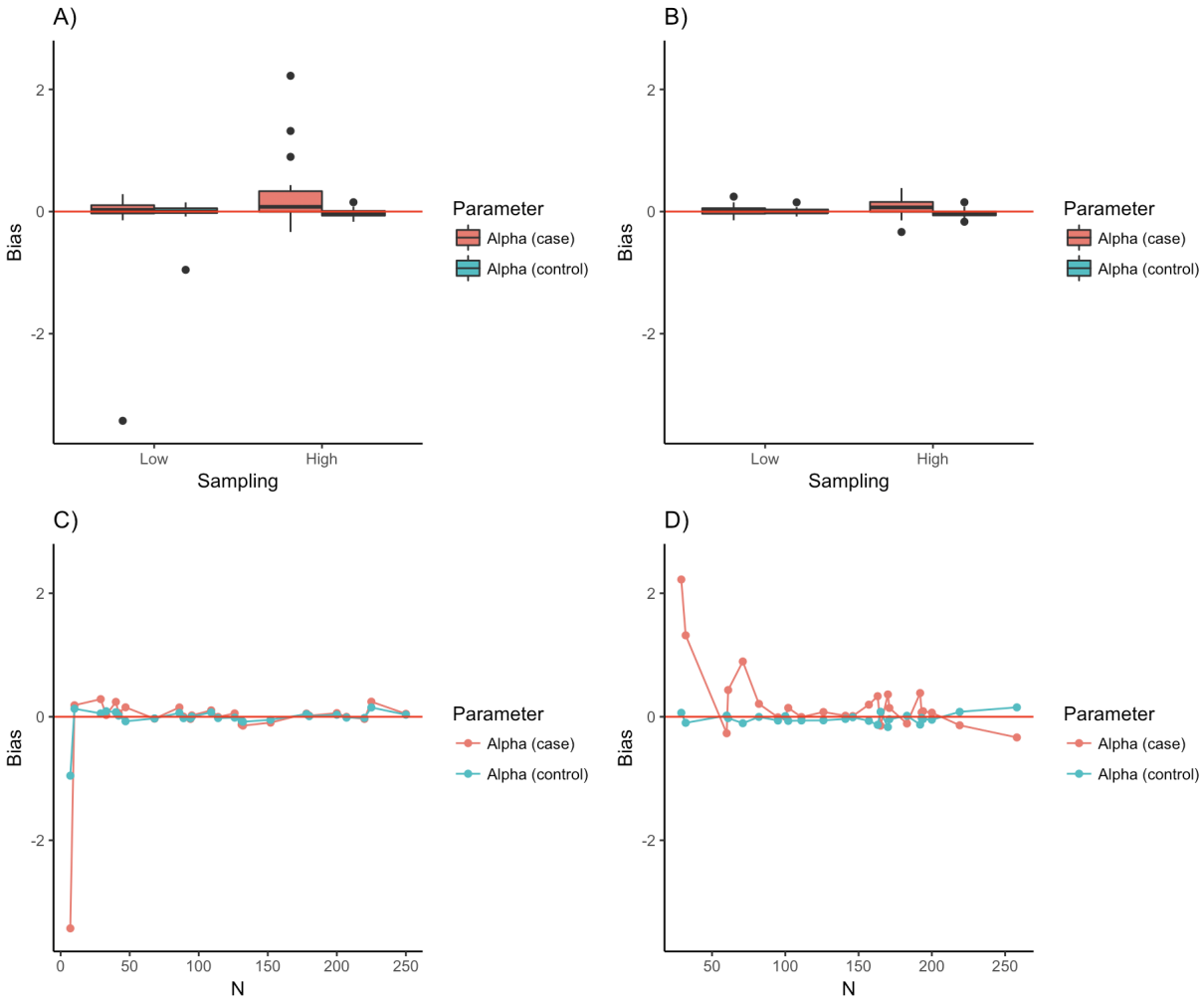

Figure 1.7: Summary of biases in preferential sampling parameters (α_+, α_-) for A) all simulated datasets and B) all simulated datasets with at least 75 observed grid cells under low and high levels of preferential sampling. C) Bias versus the number of observed grid cells under a low level of preferential sampling. D) Bias versus the number of observed grid cells under a high level of preferential sampling.

Parameters β_+ and β_- pertaining to the fixed effects $z_\lambda(x)$ of model (1.1) each consist of 3 elements, i.e. an intercept and two slope parameters, which we denote $\beta_{m,0}, \beta_{m,1}, \beta_{m,2}$ for $m \in \{+, -\}$. Under low preferential sampling average biases in $\beta_{+,0}, \beta_{+,1}$, and $\beta_{+,2}$ were -0.386, 0.078 and 0.102 respectively, while those for $\beta_{-,0}, \beta_{-,1}$, and $\beta_{-,2}$ were -0.097, 0.003, and -0.012 (Figure 1.8A). Under high preferential sampling average biases remained close to zero, but with increasing variance, both for β_+ (average biases: -0.400, 0.212, and 0.062) and β_- (average biases: -0.400, 0.212, and 0.062) (Figure 1.8B). For both low and high levels of preferential sampling, parameter biases decreased as the number of observed grid cells increased (Figure 1.8C, D), rapidly so for the case of low preferential sampling. Under high preferential sampling slight volatility in bias remained even for increasing sample sizes, especially for the intercept parameters $\beta_{+,0}$ and $\beta_{-,0}$.

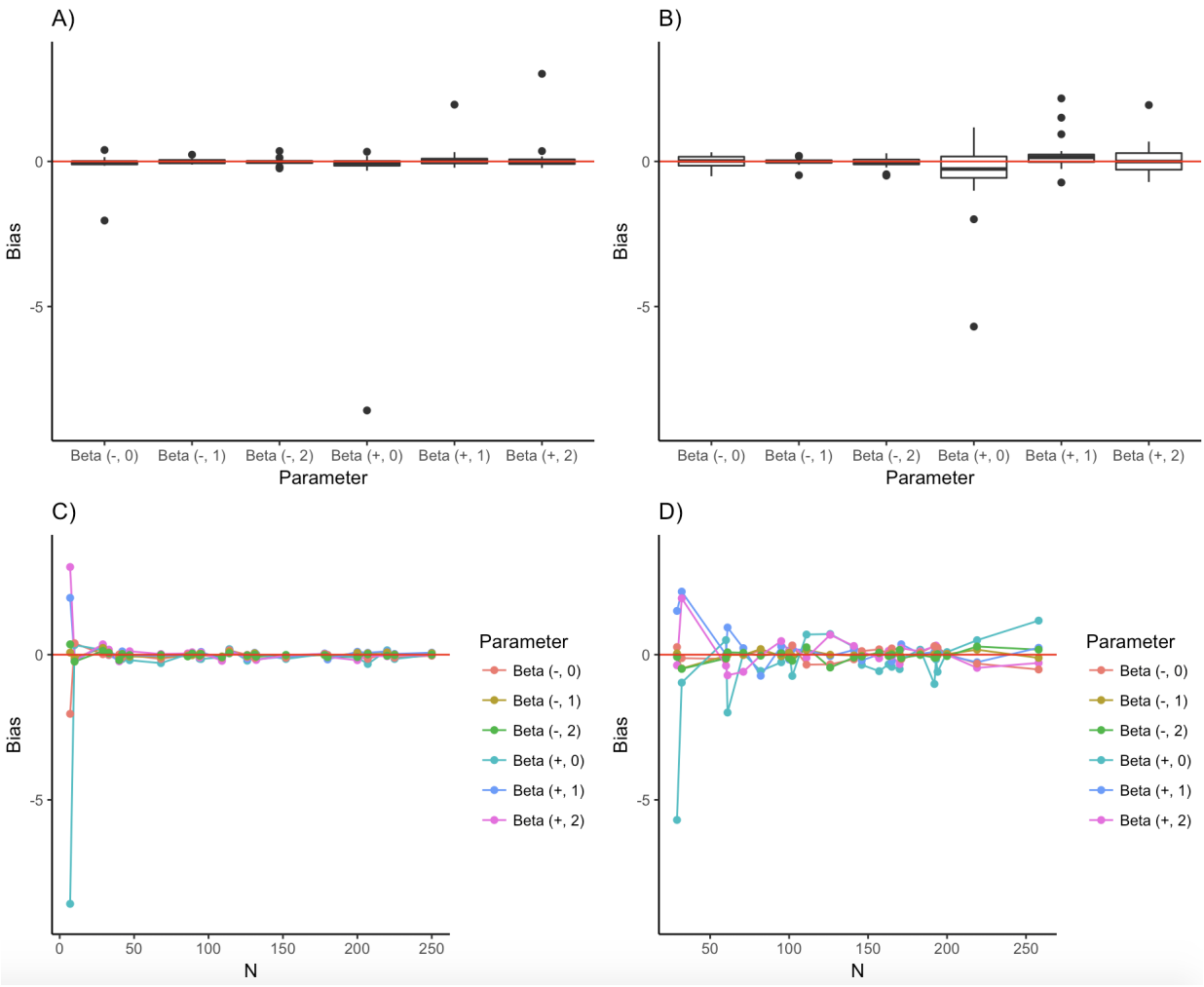

Figure 1.8: Summary of biases in estimated β_+ and β_- under A) low preferential sampling and B) high preferential sampling. Biases versus the number of observed grid cells are shown for low (C) and high (D) preferential sampling.

Lastly, the proposed model estimated the spatial range and marginal variance parameters with low bias. Under low preferential sampling, when considering only datasets with at least 75 observed grid cells, the the spatial range θ had an average bias of 0.275 (standard deviation: 1.753), while the average bias in the marginal variance ϕ was -0.511 (standard deviation: 0.652). Average biases in the range and marginal variance under high preferential sampling were 0.172 (standard deviation: 1.346) and -0.317 (standard deviation: 1.360), respectively.

1.3.5 Discussion

This simulation study demonstrated that the proposed method outperformed both reference approaches, increasingly so for higher levels of preferential sampling and sample sizes above a certain threshold. Similar to the findings of other research efforts (Diggle et al., 2010; Gelfand et al., 2012; Lee et al., 2015), this study has shown that failing to accurately model the preferential sampling mechanism can lead to biased spatial predictions, with the novelty of our study being to consider the effects of preferential sampling in the context of a disease surveillance dataset. The direction of error in our reference models (1.3) and (1.2) was also consistent with what would be expected from analyzing preferentially sampled data with traditional models which do not correct for the sampling process. That is, the tendency to overestimate disease odds naturally arises given that the data tend to be sampled in regions which are of higher disease odds. Aside from offering this reduced RMSE in estimated log disease odds, model (1.1) also estimate its individual parameters with low bias for the datasets generated in this study. Estimating α_+ and α_- with low bias is especially important due to the fact that these parameters govern the stochastic dependency between the observed case/control abundances and the distribution of observation sites.

This study also raised two significant caveats with respect to the performance of the proposed model. The first pertains to the necessary extent of spatial coverage of the study region. Specifically, performance suffered when too few grid cells of the study area were observed by the simulated surveillance system. The level of required spatial coverage varied by the level of preferential sampling, with a high level requiring roughly 75 observed cells (or 29% of all cells) for good performance and the lower level requiring fewer. In addition to the number of observed grid cells, disease prevalence also appeared to influence model performance, with considerable

error arising for prevalences below 0.05. However, we argue that these are reasonable constraints for several disease surveillance datasets, including that in the analysis chapter of this dissertation. Provided the dataset offers a coverage of at least 30% of the study region, and provided that the disease is not of very low prevalence, this study has lended support for the advantages offered by our proposed model.

1.4 Simulation 2: Parameter Initialization

1.4.1 Introduction

Markov Chain Monte Carlo, the algorithm used to sample from the posterior distribution of the novel spatial model proposed in this project, entails constructing a Markov chain designed such that the equilibrium distribution reached by the chain is the true distribution we wish to estimate. This algorithm requires specification of the initial values of the Markov chain. In practice, these values may be assigned randomly or by heuristic, such as the output of simpler statistical models. It is of crucial importance to ensure that convergence of the Markov chain is not contingent upon some unknown range of initial values.

The following simulation study probes whether such a dependence on parameter initialization arises when fitting our proposed method. We evaluate model convergence under two distinct scenarios for generating initial values: an uncalibrated manner, where initial values are assigned in an uninformative fashion, and a calibrated strategy, in which initial values are taken as parameter estimates of simpler models. We consider the model performance of these two scenarios under different prior specifications for the parameters α_m ($m \in \{+, -\}$) of model (1.1), in particular, normal and truncated normal priors. This simulation study is thus intended to ascertain the

effect of initial values on convergence of the proposed model under a variety of prior specifications.

Simulated Data

For this simulation study, a single preferentially sampled disease surveillance dataset was simulated from model (1.1). We now detail the covariates used and parameter values chosen to generate this data. The study region of California was discretized into 458 square grid cells of area $1090.3km^2$. Two spatial surfaces derived from the principal component decomposition of the PRISM climatic dataset were used as the covariates $z_\lambda(x)^T$ of model (1.1). The remaining parameters in model (1.1) were chosen to result in a disease prevalence of 19%, corresponding to 2,341 cases and 10,194 controls (Table 1.6).

Parameter	Value
α_+	1
α_-	-1
β_+	(-0.25, 0.75, -0.5)
β_-	(3, 1, 0.5)
Range	6
Marginal Variance	12

Table 1.3: Simulation parameters for the MCMC initialization study.

A total of 128 locations (Figure 1.9) of observation sites was simulated from model (1.1).

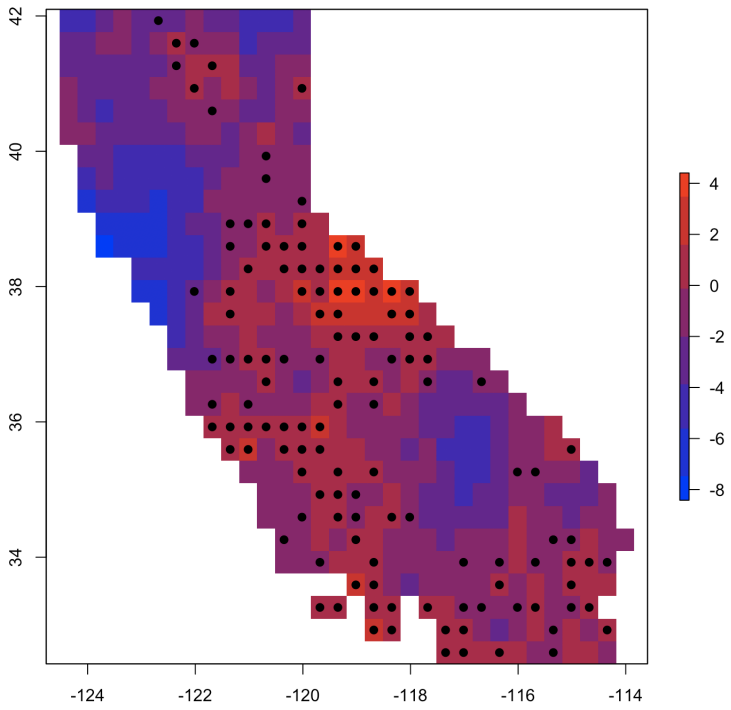

Figure 1.9: Raster of simulated observation sites for the parameter initialization study. Observed locations are denoted by black circles. Raster cells are colored according the value of realizations of the Gaussian process $w(x)$ at the centroid of each cell.

In this simulation study we evaluate the convergence of model (1.1) under a cross section of two conditions: parameter initialization and prior specification.

Parameter Initialization

We consider two strategies, calibrated and uncalibrated, for choosing initial MCMC values. The calibrated assignment uses simpler models to provide heuristic estimates for the parameters in model (1.1). The uncalibrated strategy assigns values at random, without reference to the data. We now describe both initialization strategies, beginning with the calibrated option.

Model (1.1) contains 7 parameters to fit: disease covariate related parameters β_+ and

β_-, the vector of spatial random effects w, the spatial range θ and marginal variance ϕ, as well as scalar preferential sampling parameters α_+ and α_-. In the calibrated initialization strategy, we begin by taking advantage of the fact that model (1.1) specifies the spatial pattern of observed sampling sites as a set of Bernoulli random variables with probability of success related to $w(x)$, the latent spatial process of disease risk. That is, for κ_i denoting the Bernoulli outcome of whether the ith grid cell in the study region is observed, we can obtain crude estimates for w, θ and ϕ by fitting the following spatial logistic model

$$\begin{aligned} \kappa_i|\xi(x_i) &\sim \text{Bernoulli}(\xi(x_i)) \\ \text{logit}(\xi(g)) &= w(x) \\ w(x) &\sim \mathcal{GP}(0, k(.,.;\theta,\phi)) \end{aligned} \tag{1.4}$$

where $k(.,.;\theta,\phi)$ is the exponential covariance function with range and marginal variance parameters θ and ϕ. Model (1.4) was fit to the data using an MCMC scheme in which w was updated by Hamiltonian Monte Carlo, θ by Metropolis-Hastings Random Walk, and ϕ by Gibbs sampling. Initial values of this MCMC scheme were chosen such that w was assigned to be a vector of independent standard normal random samples, θ a random sample from a Uniform(5, 7) distribution, and ϕ a random sample from a Uniform(10, 15) distribution. In this way no information from true parameter values was used to initialize the MCMC. The Markov chain was run for 1000 iterations with a burnin of 200, producing initial estimates of w, θ, and ϕ (Figure 1.10).

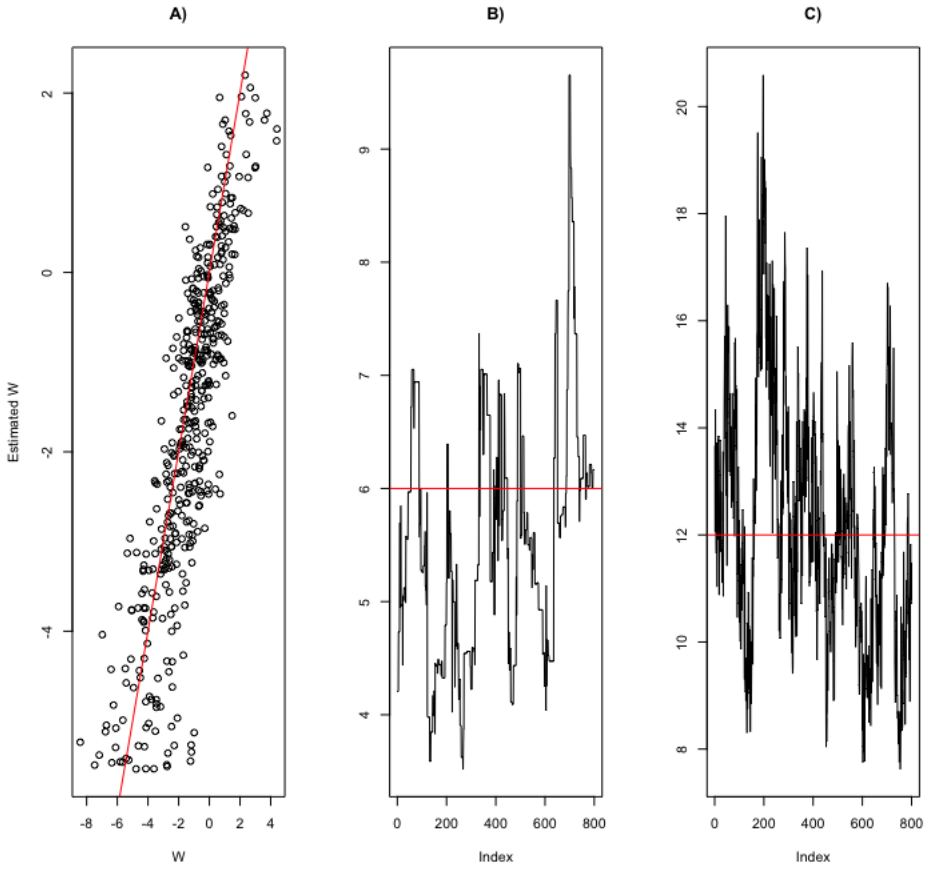

Figure 1.10: Spatial logistic regression output used for MCMC initialization. A) True versus estimated spatial random effects. Estimated effects are posterior means. B) Traceplot of the spatial range parameter θ. The horizontal red line denotes the true value of θ. C) Traceplot of the spatial marginal variance parameter ϕ. The horizontal red line denotes the true value of ϕ.

Thus, under the calibrated initialization scheme, posterior sample means of w, θ, and ϕ obtained from model (1.4) were taken to be initial values of the MCMC routine used to fit model (1.1). Given these posterior means, initial values for β_+ and α_+ were taken to be the maximum likelihood estimates obtained by fitting a Poisson regression model regressing case counts on disease covariates and the initial value of w obtained from the previous step. Similarly, β_- and α_- were initialized as estimates from a Poisson regression model with control counts as the response.

In contrast, the uncalibrated initialization strategy did rely on simpler models to

assign more informed initial values. Instead, w was initialized as a vector of zeros, β_+ and β_- as vectors of independent standard normal random variables, α_+ as a sample from Uniform(2, 3), α_- as a sample from Uniform(-3, -2), θ as a sample from Uniform(9, 10), and ϕ as a sample from Uniform(6, 8).

Prior Specifications

For each initialization scheme described in the previous section, we fit model (1.1) under two prior specifications for α_+ and α_-, i.e. the parameters representing the strength or degree of preferential sampling with regards to cases and controls, respectively. In the first specification we assume normal priors $\alpha_+ \sim N(1, 4)$ and $\alpha_- \sim N(-1, 4)$, where prior means have been chosen to equal the true values of α_+ and α_-. In the second, we assign truncated normal priors to α_+ and α_-. A truncated normal random variable has probability density function

$$f(x; \mu, \sigma, a, b) = \frac{\phi((x-\mu)/\sigma)}{\sigma(\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}$$

for $a \leq x \leq b$, where $\phi(x) = \frac{1}{2\pi}\exp(-\frac{1}{2}x^2)$ and Φ is the cumulative distribution function of the standard normal random variable. A truncated normal variable can be conceptualized as a random variable resulting from bounding the support of a normal random variable to fall between a and b. In this simulation study we assign the priors $\alpha_+ \sim \text{TruncNorm}(\mu = 1, \sigma = 2, a = 0, b = \infty)$ and $\alpha_- \sim \text{TruncNorm}(\mu = -1, \sigma = 2, a = -\infty, b = 0)$. In effect, we constrain α_+ to be positive and α_- to be negative, under the rationale that sampling in high risk locations will exert a positive influence on the abundance of cases and negative influence on that of controls.

Thus, we consider a total of 4 modeling configurations in this study, encompassing

all possible combinations of initialization strategy with prior specification of α_+ and α_-.

1.4.2 Results

Model (1.1) was fit under 4 different modeling configurations which varied MCMC initial values and prior distributions for α_+ and α_-. Under each configuration the model was fit by an MCMC implementation wherein $\alpha_+, \alpha_-, \beta_+, \beta_-$ and random effects w were updated according to separate Hamiltonian Monte Carlo samplers. Each sampler was self-tuned by dual averaging to achieve a target acceptance rate near 0.65. The spatial range parameter θ was updated by Metropolis-Hastings random walk, and the spatial marginal variance was updated by Gibbs sampling. The primary evaluation metrics of this simulation study are accuracy of the estimated log disease odds and bias in estimated parameters. The 4 configurations considered here are calibrated and uncalibrated initial MCMC values for each of normal and truncated normal prior distributions for α_+ and α_-, the parameters which reflect the strength of preferential sampling.

The truncated normal, calibrated configuration achieved the lowest root mean squared error of 1.935 in estimated log disease odds (Figure 1.11) (Table 1.4). Second was the truncated normal, uncalibrated configuration with 2.0135 RMSE, followed by the normal calibrated configuration at 2.681. The normal uncalibrated setting had the highest RMSE at 5.569.

Configuration	RMSE
Normal uncalibrated	5.569
Normal calibrated	2.681
Truncated normal uncalibrated	2.015
Truncated normal calibrated	1.935

Table 1.4: Root mean squared error in estimated log disease odds of 4 different MCMC configurations. The configurations considered are the combination of normal versus truncated normal priors for α_+ and α_- with calibrated versus uncalibrated initial MCMC values.

The distribution of errors in estimated log disease odds also differed greatly by MCMC configuration (Figure 1.11). The normal, uncalibrated strategy tended to overestimate lower log odds and underestimate higher values. The truncated normal, uncalibrated configuration also underestimated lower log odds, but to a much lesser extent. Errors were more evenly distributed about the true values for the normal, calibrated and truncated normal, calibrated configurations, but with error tending to increase as true log odds decreased.

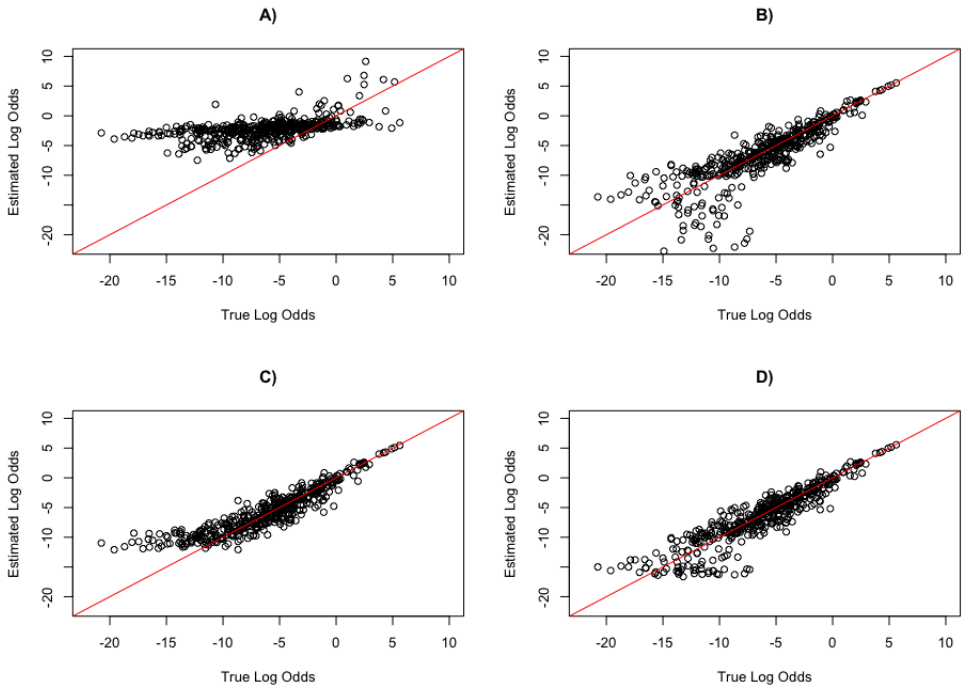

Figure 1.11: Estimated log disease odds under different MCMC configurations. A) Uncalibrated initial MCMC values with normal priors for α_+ and α_-. B) Calibrated initial MCMC values with normal priors for α_+ and α_-. C) Uncalibrated initial MCMC values with truncated normal priors for α_+ and α_-. D) Calibrated initial MCMC values with truncated normal priors for α_+ and α_-.

Model	Parameter	Estimate	True Value	Bias
Normal uncalibrated	α_+	1.463	1	0.463
Normal calibrated	α_+	0.999	1	-0.001
Truncated normal uncalibrated	α_+	1.288	1	0.288
Truncated normal calibrated	α_+	1.13	1	0.13
Normal uncalibrated	α_-	1.63	-1	2.63
Normal calibrated	α_-	-0.971	-1	0.029
Truncated normal uncalibrated	α_-	-1.2	-1	-0.2
Truncated normal calibrated	α_-	-1.096	-1	-0.096

Table 1.5: Preferential sampling parameter estimates under different MCMC configurations.

MCMC convergence for each configuration was assessed by traceplots. The number of MCMC samples drawn and length of burnin periods varied sizably depending on parameter initialization strategy and prior specification (Table 1.6), with the truncated normal models having the smallest chain length of 4000 and the normal uncalibrated model with the greatest length of 14000.

Configuration	Samples	Burnin
Truncated normal calibrated	4,000	500
Truncated normal uncalibrated	4,000	1,500
Normal calibrated	2,000	500
Normal uncalibrated	14,000	6,000

Table 1.6: Number of MCMC samples and burnin periods for different prior specifications and parameter initialization strategies.

1.4.3 Discussion

The objective of this simulation study is the assess convergence of the proposed preferential sampling model (1.1) under different prior distributions and MCMC initializations. Recall that we consider both calibrated and uncalibrated strategies for setting initial MCMC values, with calibrated initialization taking the form of outputs of simpler models, and uncalibrated initialization simply assigning random initial states. In either case initial values are not informed by knowledge of the true parameters, a crucial requirement for the applicability of model (1.1) to real world scenarios. For each initialization strategy, we fit model (1.1) under normal and truncated normal prior distributions for both preferential sampling parameters, α_+ and α_-. We refer to the 4 MCMC configurations resulting from these different initializations and priors as: 1) normal, uncalibrated 2) normal, calibrated 3) truncated normal, uncalibrated, and 4) truncated normal, calibrated.

The truncated normal, calibrated configuration achieved the best root mean squared

error in predicted log disease odds, followed closely by the truncated normal, uncalibrated model (Figure 1.11) (Table 1.4). The normal, calibrated configuration achieved the next lowest RMSE, trailing the preceding two configurations by a modest but not exorbitant margin. However, the normal, uncalibrated strategy had by far the highest RMSE of any configuration.

All four strategies showed an increasing error in predicted disease log odds as the true log odds value decreased (Figure 1.11). This error was roughly normally distributed about the true value for the normal uncalibrated version, slightly positively skewed for the truncated normal, uncalibrated model, slightly negatively skewed for the truncated normal calibrated version, and greatly positively skewed for the normal uncalibrated configuration. Consequently, the uncalibrated normal model resulted in a substantially higher RMSE than the three contrasting configurations, with a value of 5.568, over twice the value of the next worst result of 2.681 for the normal, calibrated model.

The normal uncalibrated model not only fails to accurately predict log disease odds but also suffers considerable bias in parameter estimation, due to failure of convergence. The MCMC routine under this configuration failed to converge to proper values of the preferential sampling parameters α_-, and showed moderate bias in α_+ (Table 1.5). In fact, the estimated value of 1.63 is completely of the wrong sign, with the true value of α_- falling at -1. In contrast, the other three MCMC configurations estimate α_+ and α_- within reasonable levels of bias (Table 1.5). Moreover, the failure of the normal uncalibrated model to converge to the proper value of α_- is coupled with substantial bias in β_+ and β_-, Thus, the normal uncalibrated model not only fails to provide quality estimates of disease risk but also suffers considerable bias in parameter estimation, whereas the other models perform reasonably well in both risk prediction and parameter estimation.

This simulation study has yielded key insights regarding the conditions under which the proposed preferential sampling model performs well, showing that model performance depends on both MCMC initialization and prior specification. Model performance, as measured by both accuracy in predicted disease risk and statistical bias in parameter estimation, was greatest for the truncated normal, calibrated modeling configuration. The next best strategy is arguably the normal, calibrated strategy, which achieves a similar accuracy to the truncated normal calibrated model but results in moderately lower bias in parameter estimates. Lastly, and perhaps most crucially, we have shown that the normal uncalibrated scheme fails to deliver quality risk predictions and parameter estimates.

This simulation study has shown that using calibrated initial MCMC values, i.e., values obtained from fitting simpler models to the data, is a key ingredient to model success when truncated normal priors are not assigned to α_+ and α_-, which may not always be appropriate. One drawback of the calibration process is the fact that a spatial logistic regression model must be fit, demanding extra time and computational resources. Within the specific context of this simulation this additional model fitting is largely unproblematic, as convergence is quickly obtained. However, at substantially higher resolutions the additional modeling step may become more burdensome. Additionally the success of the calibration step hinges on quality estimates of the spatial random effects w from locational data, which may be infeasible to obtain if there are too few observed locations. In such cases where model calibration fails the truncated normal, uncalibrated configuration may suffice.

The appropriateness of the truncated normal prior specification for α_+ and α_- ultimately rests on the assumption that preferential sampling tends to assign observation sites in areas with a greater abundance of cases and lesser abundance of controls. The truncated normal prior distribution does assign non-zero probability to $\alpha = 0$, i.e.

a non-preferential sampling scheme, which is crucial to the real world applicability of the model given the possibility that datasets thought to be preferentially sampled may in fact possess no or a merely weak relationship between sampling locations and disease risk. However the lack of support for $\alpha_+ < 0$ may result in positively biased estimates of α_+ when the true value of α_+ is low, while the same may hold for α_- but in the opposite direction. But more importantly, it may be inaccurate to assume that preferential sampling should be not be associated with an increased abundance of controls, insofar as the increase in controls is matched with an even greater increase in the abundance of cases so that the overall risk is higher in sampled locations. In such instances both α_+ and α_- may be greater than zero, under the condition that $\alpha_+ > \alpha_-$. For example, scenarios may arise where high risk regions for a disease tend to have greater numbers of both cases and controls compared with low risk areas, but the increase in the number of cases is greater than that of controls, resulting in an overall greater risk. For this reason it may be unreasonable to assume $\alpha_- < 0$ in every real world sampling application. Thus, our ultimate recommendation for fitting model (1.1) is the normal prior specification with calibrated initial values. In situations where there is clear evidence that the assumption of $\alpha_- < 0$ is valid, such as those where previous studies show a decreased abundance of controls in high risk areas, then we would recommend the truncated normal, calibrated scheme, followed by the truncated normal, uncalibrated option. We do not recommend the normal, uncalibrated scheme under any circumstance. Future simulation studies and analyses in this dissertation adopt the normal, calibrated model fitting strategy to allow for the most realistic flexibility in the range of α_- while still preserving quality of model performance.

1.5 Analysis

1.5.1 Introduction

This analysis applies the proposed method to a disease surveillance dataset obtained from the California Department of Public Health (CDPH). The target of surveillance is *Yersinia pestis* infection, otherwise known as plague, in the rodent family of squirrels (*Sciuridae*) within the state of California. The surveillance system collects data by means of traps laid out to capture rodents at specific locations, from which recovered rodents are tested for plague. Due to limited resources and managerial objectives, CDPH tends to assign these sampling locations to areas at high suspected risk for plague, making the data collection mechanism a case of preferential sampling. Our objective is thus to determine whether modeling the sampling process can offer an improved plague risk map over the state of California, or at least, a risk map which is quantitatively distinct from that obtained by alternative methods which make no attempt to address preferential sampling.

The CDPH surveillance system targets plague in the rodent family of squirrels, known as *Sciuridae* or *Sciurids*. A total of 21 different species within this family have been recovered by surveillance, namely the: Antelope Ground Squirrel, Antelope Ground Squirrel (WhiteTail), Belding's Ground Squirrel, California Ground Squirrel, Chipmunk, Least Chipmunk, Long-eared Chipmunk, Lodgepole Chipmunk, Merriam's Chipmunk, Panamint Chipmunk, Shadow Chipmunk, Siskiyou Chipmunk, Sonoma Chipmunk, Uinta Chipmunk, Yellow-pine Chipmunk, Golden-mantled Ground Squirrel, Ground Squirrel, Yellow-bellied Marmot, Pine Squirrel, and Squirrel. The surveillance system collects data by conducting a series of sampling events at locations throughout California. For each sampling event, Sciurids are trapped and subsequently tested for *Yersinia Pestis*. The data contain samples collected between 1983

and 2015. This analysis aggregates data for all Sciurid species, and for all years observed. Analyses by species and by specific time intervals are included in future chapters of this dissertation, with the intent for this initial analysis being to establish whether there is an observable impact of preferential sampling on the temporally aggregated data.

The surveillance system predominantly assigns sampling locations to high risk or high impact areas, where risk is assessed to be high in what are viewed as plague endemic regions, as determined by by historic cases of plague in humans or recovered Sciurid specimen, and high impact areas are regions where cases of plague in humans would be particularly damaging, such as in national parks. This sampling strategy fits the mold of preferential sampling, given that sampling locations are assigned to areas thought to be of high value for the response (disease risk) being measured. While sampling at high risk locations is often sensible from a managerial perspective, the downside is that it may result in a statistically biased estimation of the underlying risk surface (Diggle et al., 2010; Lee et al., 2011; Gelfand et al., 2012; Lee et al., 2015). Ultimately, there are two contrasting aims in disease surveillance that come in conflict here: 1) to monitor most sensitive areas in order to respond quickly to threats, all the while using constrained resources, and 2) to estimate an unbiased risk surface over a broad extent. Preferential sampling lends itself to the former, while possibly impacting the ability to perform the latter. Thus our goal is to propose a model that can estimate an improved, less statistically biased, risk surface from preferentially sampled data.

We compare our proposed model against two alternatives. The first reference model (1.3) regresses case and control counts through independent Poisson models whose rates are log-linear in climate related covariates. For a study region discretized into $g = 1, \ldots, G$ grid cells, we suppose that $i = 1, \ldots, K$ of these have been observed

by the surveillance system. We then let Y_{im} denote the count of cases ($m = +$) or controls ($m = -$) for the ith observed cell, and suppose that Y_{im} are Poisson distributed with rates $\lambda_m(x_i)$ for $x_i \in \mathbb{R}^2$

$$\begin{aligned} Y_{im} &\sim \text{Poisson}(\lambda_m(x_i)) \\ \lambda_m(x_i) &= \exp(z_\lambda(x_i)^T \beta_m) \end{aligned} \tag{1.5}$$

Inclusion of the i subscript in our notation for $z_\lambda(x_i)$ is intended to convey that the rate λ_m for the ith grid cell is a function of the covariate values at the centroid of the ith grid cell, denoted x_i. For this rate construction to hold we must assume that the grid cells are of sufficiently small area such that the values of the spatial covariates $z_\lambda(x)$ are constant throughout any given cell.

The next reference model (1.2) follows a similar design to the first, except that it includes spatial random effects realized from a Gaussian process in the structural components of the Poisson models for cases and controls. The inclusion of spatial random effects $w_+(x)$ and $w_-(x)$ is intended to capture unexplained spatial variation in case and control counts, respectively. Both Gaussian processes are modeled with exponential covariance functions, $k(x, x'; \theta_m, \phi_m)$, where spatial range ($\theta_m$) and marginal variance (ϕ_m) parameters differ across cases and controls to allow a more flexible model.

$$
\begin{aligned}
Y_{im} &\sim \text{Poisson}(\lambda_m(g_i)) \\
\lambda_m(x) &= \exp(z_\lambda(x)^T \beta_m + w_m(x)) \\
w_m(x) &\sim \mathcal{GP}(0, k(.,.;\theta_m, \phi_m))
\end{aligned}
\tag{1.6}
$$

Similarly to our proposed model, this spatial Poisson model is also fit at a lower resolution (422 km^2) due to the inefficiency of inverting the covariance matrices for both Gaussian processes $w_+(x)$ and $w_-(x)$. Estimated spatial random effects obtained from the low resolution model fit are then downscaled to a resolution of $16km^2$ by thin plate spline interpolation, allowing for the calculation of high resolution disease log odds and risk.

1.5.2 Data

Models (1.1), (1.5) and (1.6) were fit to the surveillance data over all recorded years, 1983 - 2015, for all Sciurid species recovered by the surveillance system. A total of 1,401 plague positive specimen were present in this dataset, along with 20,366 plague negative rodents for an overall prevalence of 6.43%.

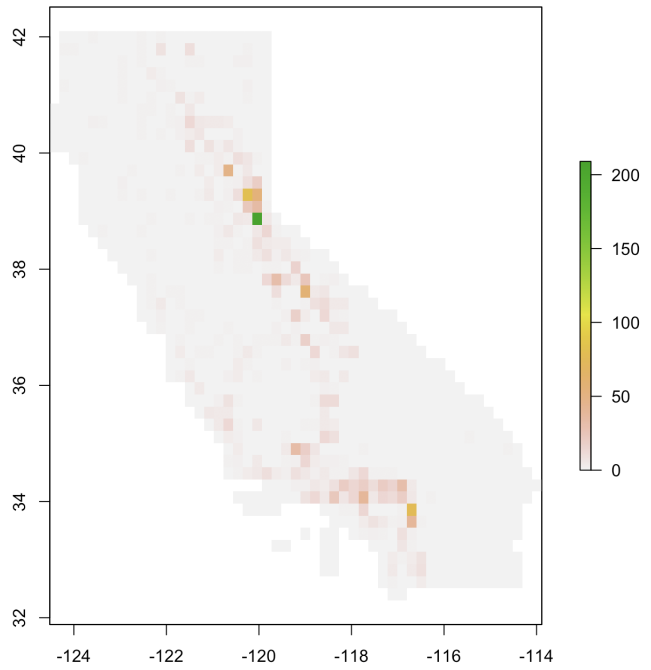

Figure 1.12: Distribution of Sciurid sampling locations between 1983 and 2015 at a resolution of 426 km^2. Grid cells are colored according to the number of distinct sampling locations therein.

Information derived from the PRISM climatic dataset, maintained by the Oregon State University, was used as the basis for disease related spatial covariates $z_\lambda(x)$. The PRISM data used here consist of a variety climatic measurements conducted at the 16 km^2 resolution, known as 30 year average normals. Specifically, mean temperature, maximum temperature, minimum temperature, precipitation, minimum vapor pressure deficit, maximum vapor pressure deficit, and mean dew point temperature were considered. However, these values were not directly used as the spatial covariates $z_\lambda(x)$, but rather, were first range standardized and then dimensionally reduced by principal component analysis. The range standardization was calculated as

$$r_{ik} = (x_{ik} - \min_i)/(\max_i - \min_i) \text{ for } (i = 1, \ldots, 7)$$

where i indexes the measurement type, k indexes the raster cell in the study region for

which the measurement was taken, $\min_i$ is the minimum value of the ith measurement and $\max_i$ is the maximum value of the ith measurement. Principal components of the 7 range standardized measurements were calculated using the rasterPCA function of the RStoolbox package, from the R programming language. The first 2 principal components (Figure 1.13) were then range standardized and scaled transformed before use as covariates $z_\lambda(x)$. An intercept term was also included in $z_\lambda(x)$. The first principal component corresponds primarily to the temperature related variables, while the second is comprised mostly of moisture related variables.

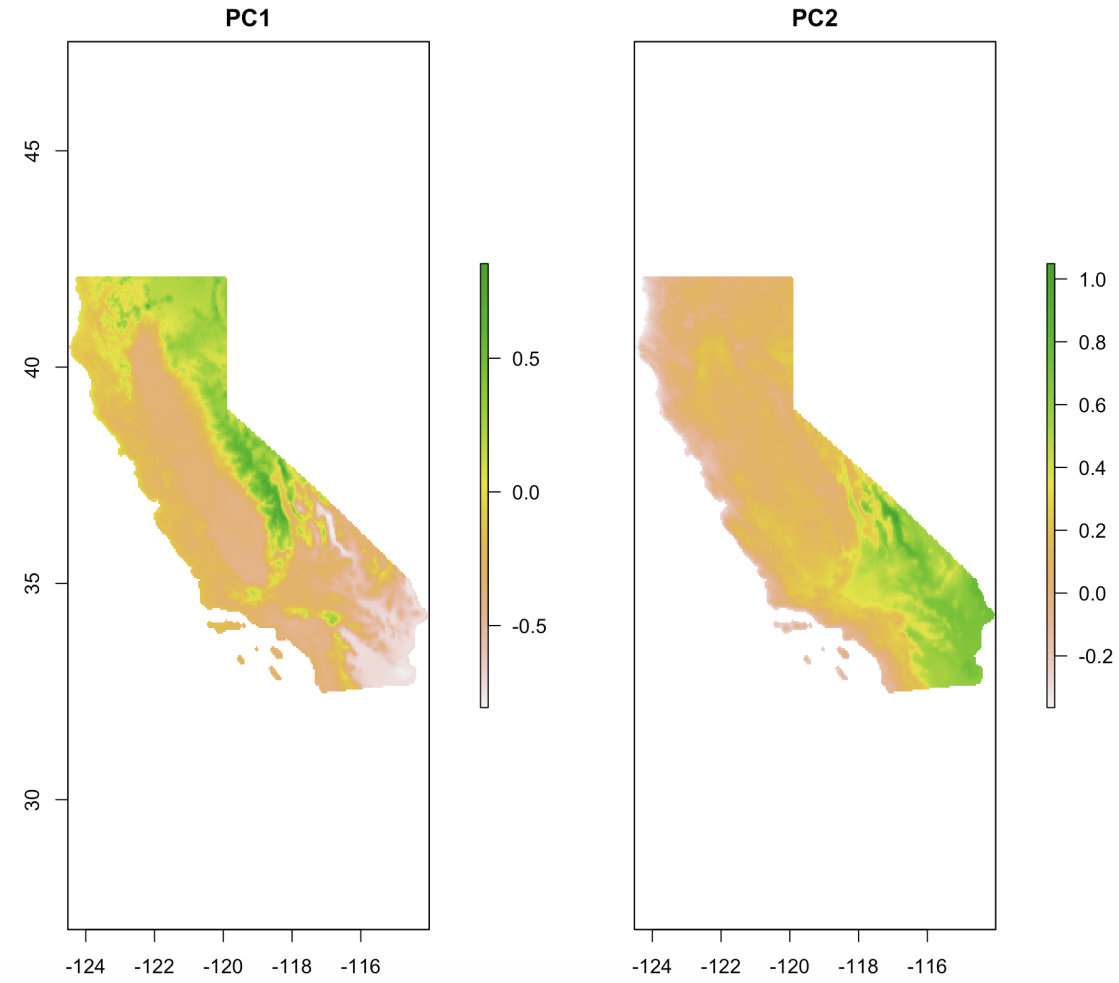

Figure 1.13: The first two PRISM principal components used as covariates for CDPH Sciurid analysis. Each raster is at a 16 km^2 resolution.

1.5.3 Results

The proposed model (1.1) was fit to the data by MCMC with a total of 10,000 samples and a burnin period of 2,000, at a resolution of 426 km^2, which discretized the study region into 1,117 grid cells. The results of model fitting were subsequently spatially downscaled to a resolution of 16 km^2 (Figure 1.14).

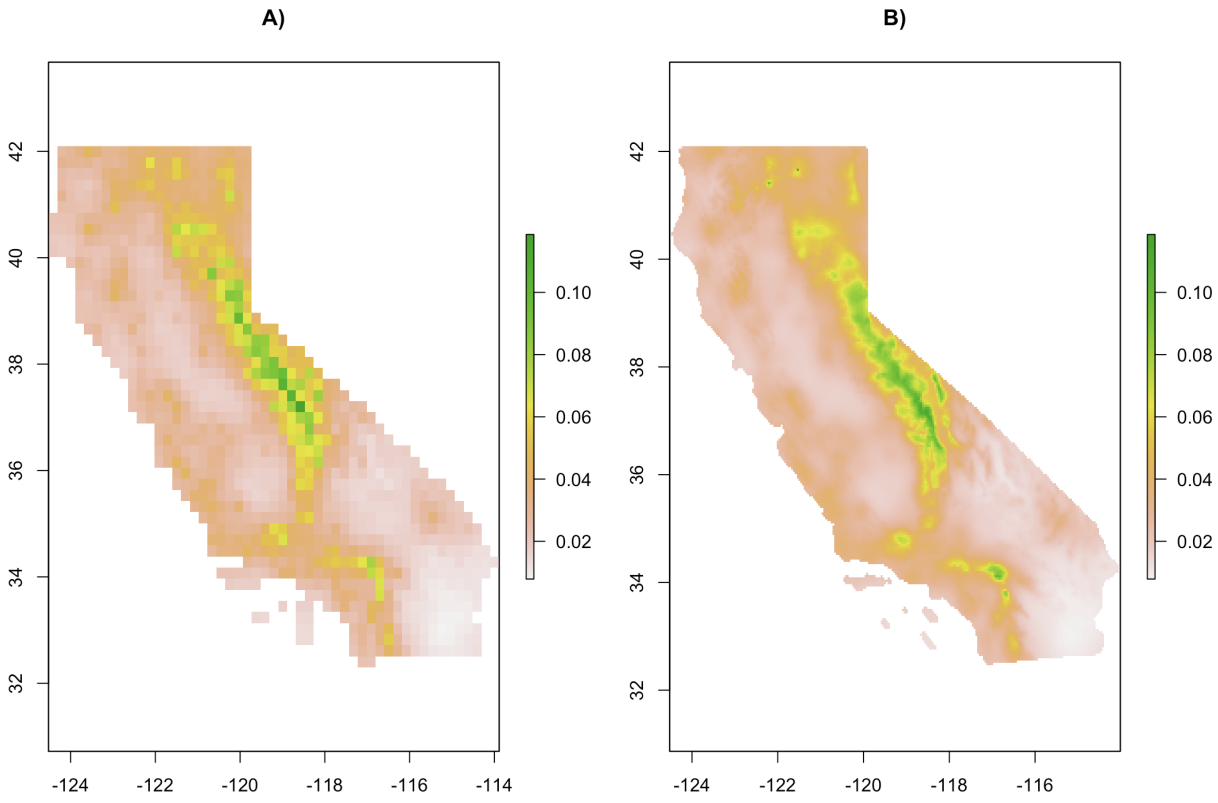

Figure 1.14: A) 426 km^2 Sciurid plague risk map and B) Spatially downscaled 16 km^2 Sciurid plague risk map.

The high resolution risk map estimated by the proposed model (1.1) is presented in Figure (1.15), and again in Figure (1.16) with the inclusion of county lines. In these risk maps the raster values represent the probability a rodent sampled at a particular location will be plague positive. Estimated risk ranges in value from 0.008 to 0.115 over the study region. Peak areas of risk fall along the Sierra Nevada mountain range,

stretching diagonally from roughly the 40th to 35th latitude towards the eastern border of the state, as well as thin pockets of elevated risk in the northeastern portion of the state, in addition to circular regions of elevated risk towards the southwestern part of the map. The southern and central coastlines also show mild elevation in risk relative to some of the lower risk regions of the map, such as the San Joaquin Valley, to the west of the Sierra Nevada mountains, and the Imperial Valley region, in the southeastern corner.

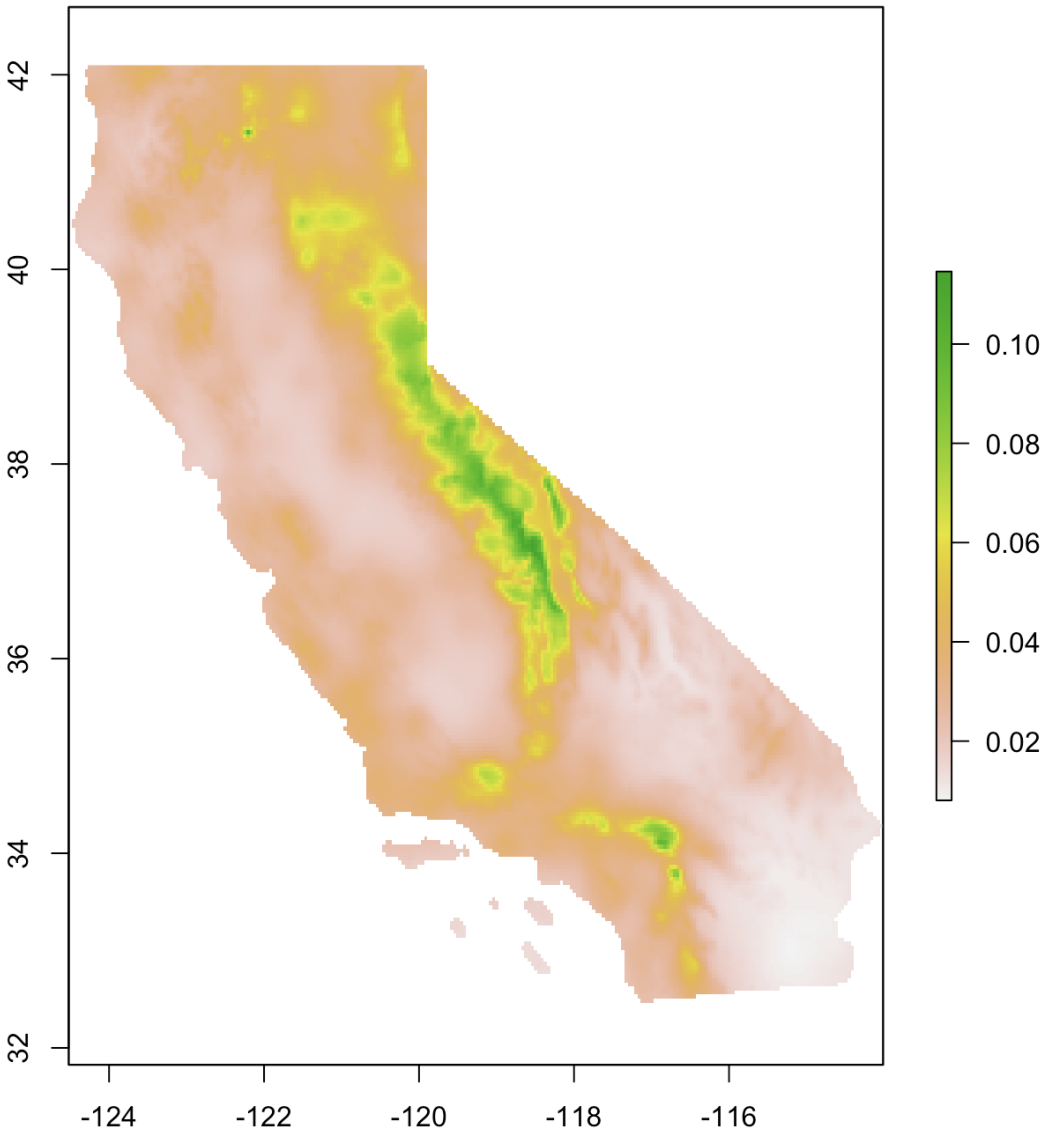

Figure 1.15: Sciurid plague risk map over California at high resolution (16 km^2), as calculated by model (1.1).

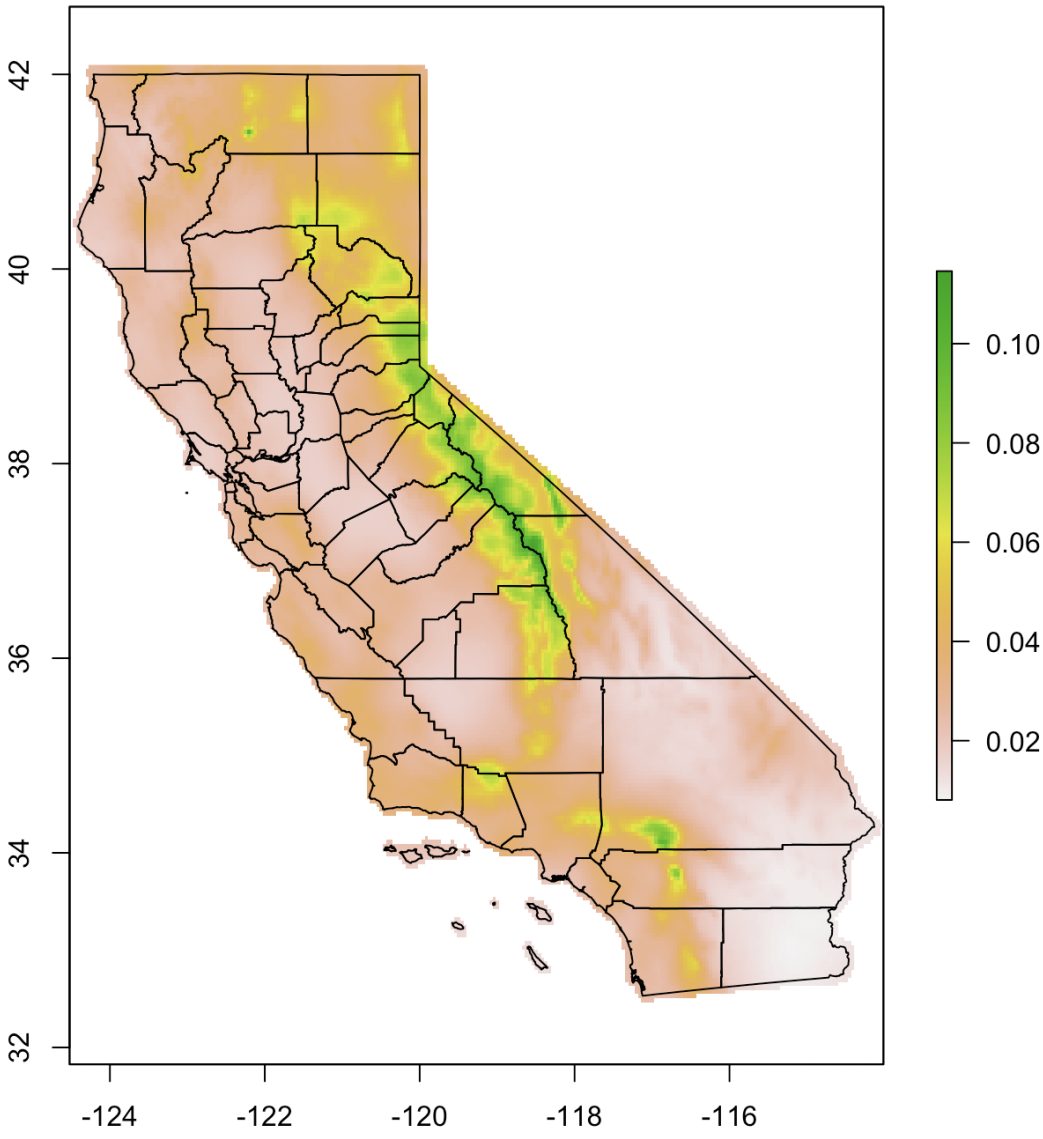

Figure 1.16: Sciurid plague risk map over California at high resolution (16 km^2), as calculated by model (1.1), with county lines.

For a rough visual inspection of the correspondence between the distributions of observed cases and controls with the resulting risk estimates, we overlay juxtapose the estimated risk map in alongside rasters of observed case and control counts (Figure 1.17). It is apparent that the elevated band in risk running along the Sierra Nevada mountain range, between the 40th and 36th parallels, coincides with a strong presence of recovered cases. In addition, the neighborhoods of increased risk in the southwestern portion of the map also contain an abundance of cases. Of note is the fact that the northeastern most pocket of elevated risk, between roughly the 41st and 42nd latitude, near the eastern state boundary, evidences plague negative specimens but not any recovered cases.

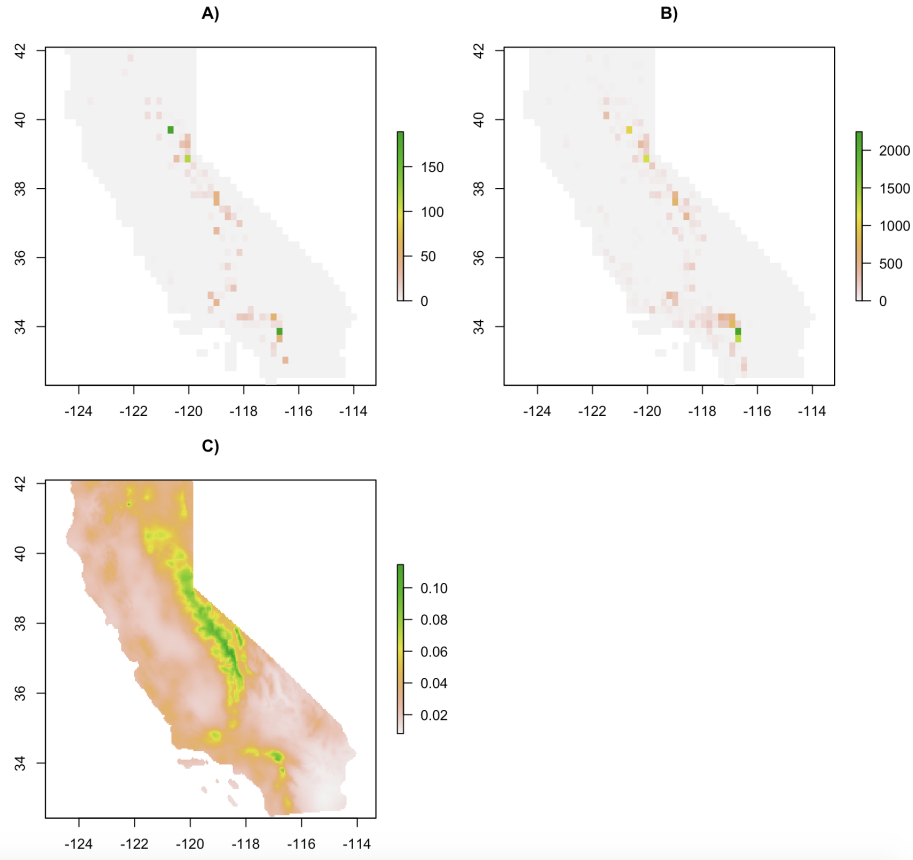

Figure 1.17: Overview of A) case counts, B) control counts, and C) the estimated risk surface for plague in Sciurids between 1982 and 2015.

This last observation, i.e. the small neighborhood of increased risk in the northeastern

corner, begs the question of why risk can be heightened in places where no cases have been recovered. To help explain this question and the mechanics behind other areas of elevated risk in the Sciurid plague map, we separate the contributions to predicted risk put forward by covariates and random effects. We recall from model (1.1) that the log disease odds for a particular point in space x are given by the difference of case and control log rates

$$z_\lambda(x)^T\beta_+ + \alpha_+ \times w(x) - z_\lambda(x)^T\beta_- + \alpha_- \times w(x)$$

We can separate this expression into firstly a covariate contribution, $z_\lambda(x)^T\beta_+ - z_\lambda(x)^T\beta_-$, due to the PRISM climatic principal components used here as fixed effects $z_\lambda(x)^T$, and secondly the expression $\alpha_+ \times w_s(x) - \alpha_- \times w(x)$ provided by the spatially structured random effects $w(x)$. Side by side inspection of these different contributions helps explain the areas of high risk as either due to covariates, or random effects, or a mixture of both (Figure 1.18). We see that the northeastern pocket of elevated risk has high levels of covariate contributions, with a much fainter contribution from random effects, suggesting that the risk increase here is primarily covariate driven. Turning our attention toward other areas of the map, we observe the high risk region along the Sierra Nevada mountains, between the 37th and 35th parallels, to be underpinned by both high covariate and random effect contributions.

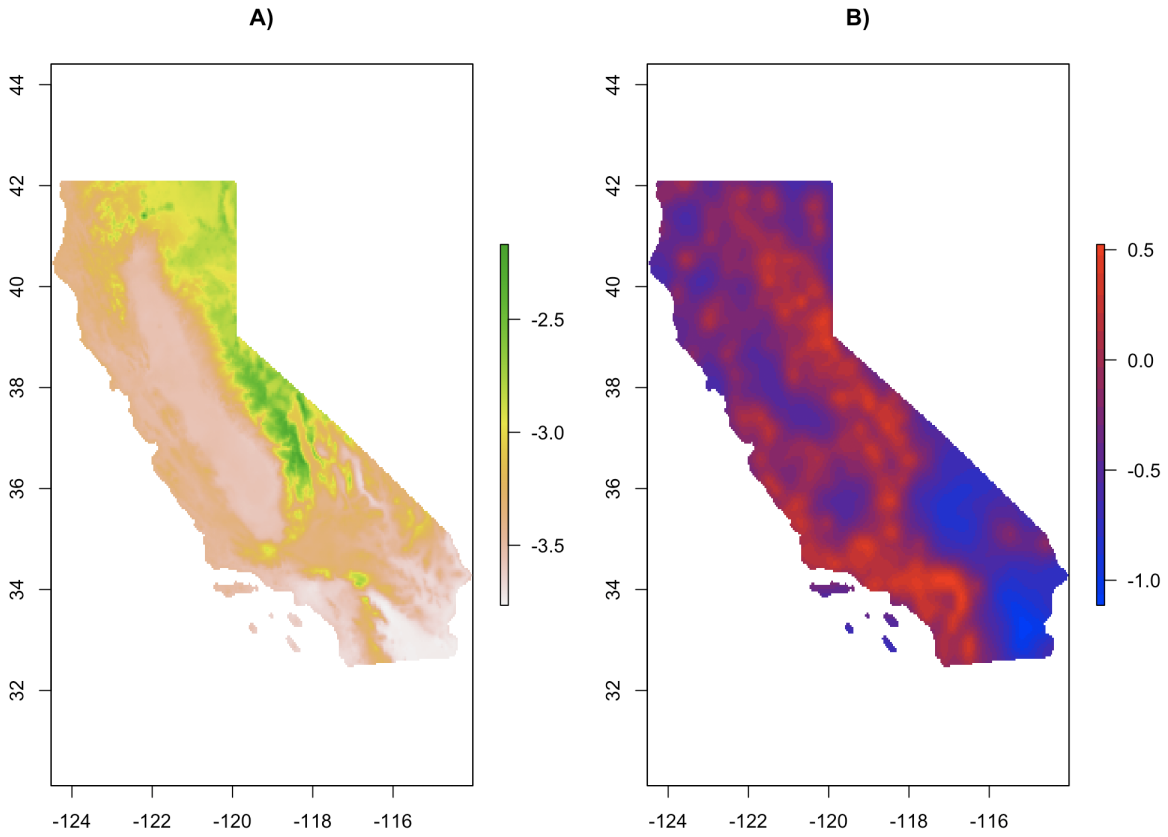

Figure 1.18: Comparison of A) covariate versus B) random effect influence on estimated log disease odds of plague in Sciurids, showing A) the value contributed to the log odds by covariate effects, $z_\lambda(x)^T\beta_+ - z_\lambda(x)^T\beta_-$ and B) that contributed by random effects, $(\alpha_+ \times w) - (\alpha_- \times w)$.

The raster of posterior variances in predicted risk, obtained from the resolution at which model (1.1) was fit before downscaling, shows an overall low level of posterior variance, ranging from 3.074×10^{-6} to 1.614×10^{-4}, with a median value of 2.494×10^{-5}. A slight east-west gradient of decreasing variance is apparent, with increased posterior variance arising along the eastern border of the state.

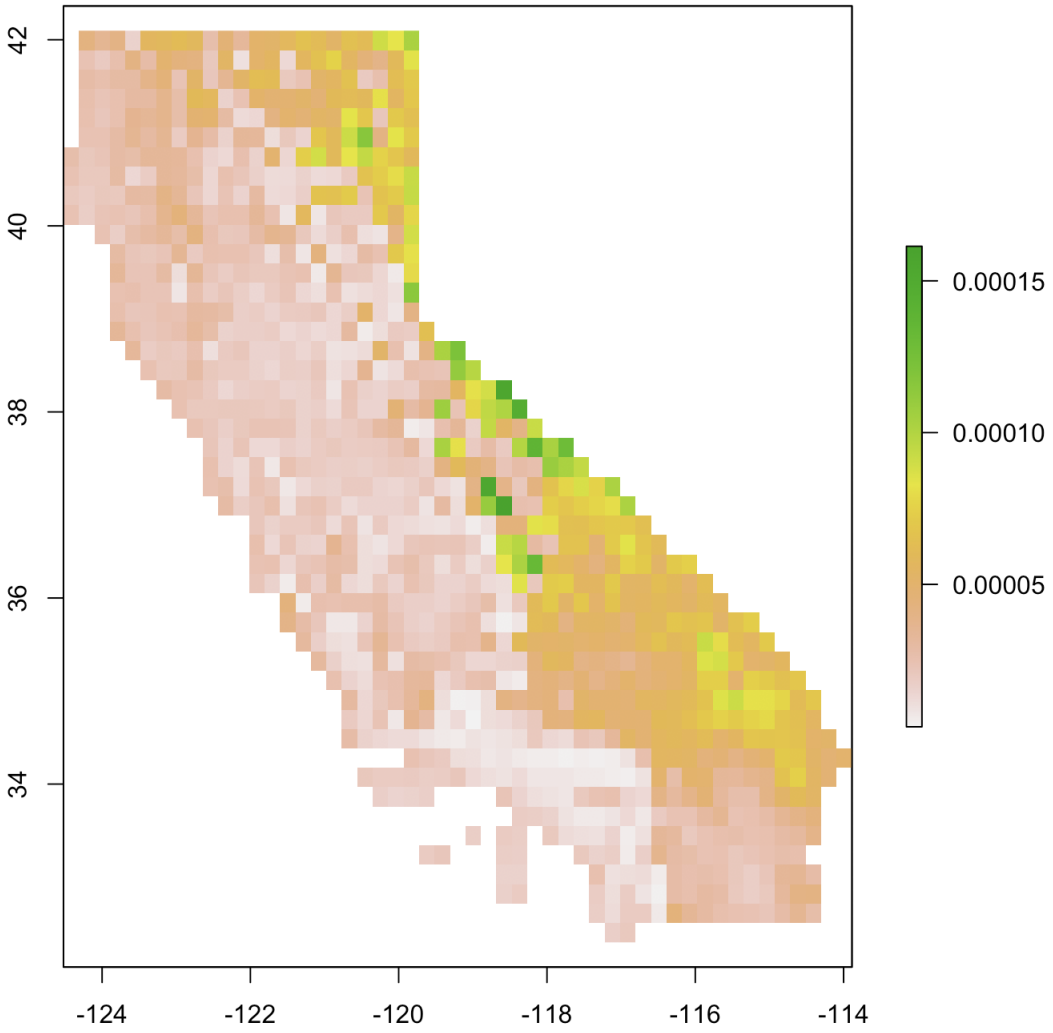

Figure 1.19: Posterior variance in predicted risk for plague.

Stark differences in predicted risk arise between our proposed method and the two reference approaches (Figure 1.20). The Poisson model shows a generally higher level of estimated risk compared to model (1.1), particularly in the northern area of the state, but also along those regions which were predicted to be of very low risk by model (1.1), namely the San Joaquin and Imperial valleys.

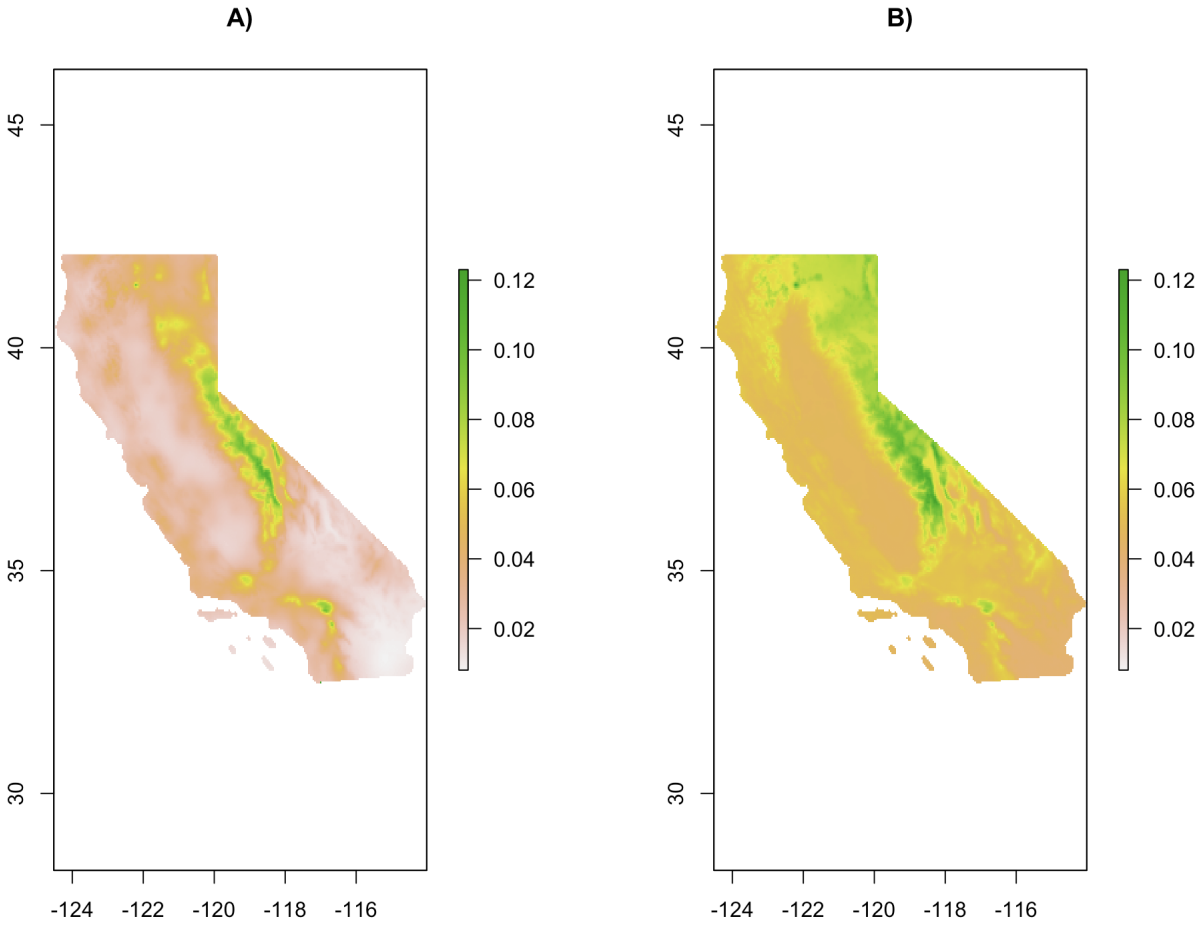

Figure 1.20: Comparison of Sciurid plague risk maps over California at 16 km^2 resolution. A) Risk map adjusted for preferential sampling by model (1.1). B) Unadjusted risk map obtained from the reference Poisson regression model (1.5).

While the Poisson model shows notable differences in predicted risk from the proposed method, the spatial Poisson model (1.6) diverges to an even greater degree (Figure 1.21). Model (1.6) predicts several pockets of substantially high risk, reach-

ing a maximum value of 0.59, outside of which the predicted risk remains very low. These areas of elevated risk fall primarily along the Sierra Nevada mountain range, in addition to key sections of Southern and Northern California.

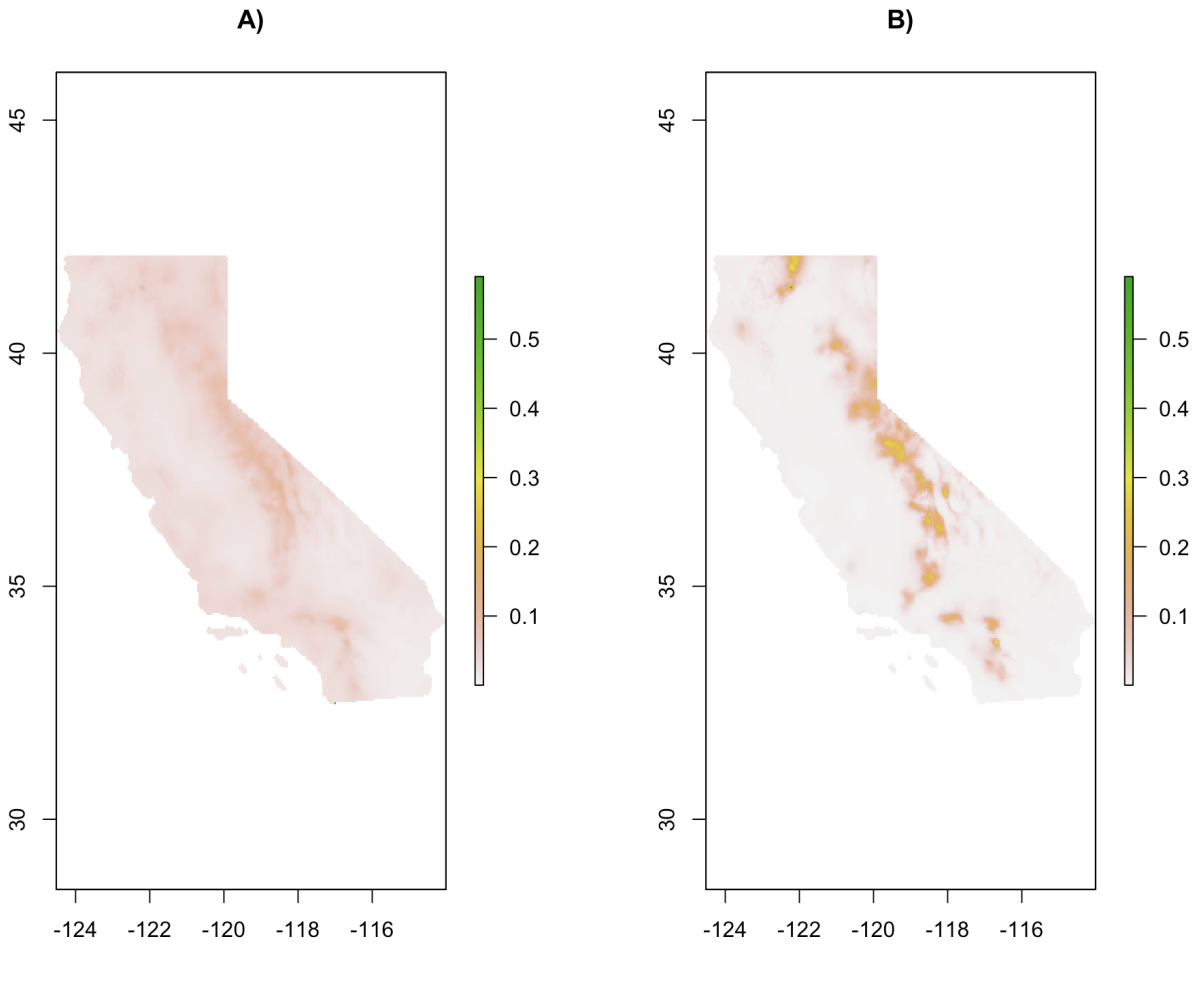

Figure 1.21: Comparison of Sciurid plague risk maps over California at 16 km^2 resolution. A) Unadjusted risk map obtained from the reference spatial Poisson regression model (1.6). B) Risk map adjusted for preferential sampling by model (1.1).

In addition to these visual comparisons, we also examine the cell by cell differences in estimated risk between models (1.1) and (1.5) with a scatterplot (Figure 1.22 A), where the (x, y) coordinates of each point represent the corresponding risk calculated by models (1.5) and (1.1), respectively, for each grid cell in the study region, and the red diagonal is a line of slope 1 and intercept 0. Consequently, points falling

below the red line indicate higher predicted values for the Poisson model than the proposed method, and points falling above the line signify the opposite. Here, the reference model generally tends to overestimate risk relative to model (1.1) for the vast majority of grid cells, while still underestimating some values relative to model (1.1). We also compare the cell by cell estimated risk differences between model (1.1) and the spatial Poisson model (1.6) (Figure 1.22 B). The divergence in cell by cell risk between these two models is far more pronounced than that between model (1.1) and the Poisson model, with several values falling substantially lower for the proposed method, but also a large number of values estimated to be higher by model (1.1).

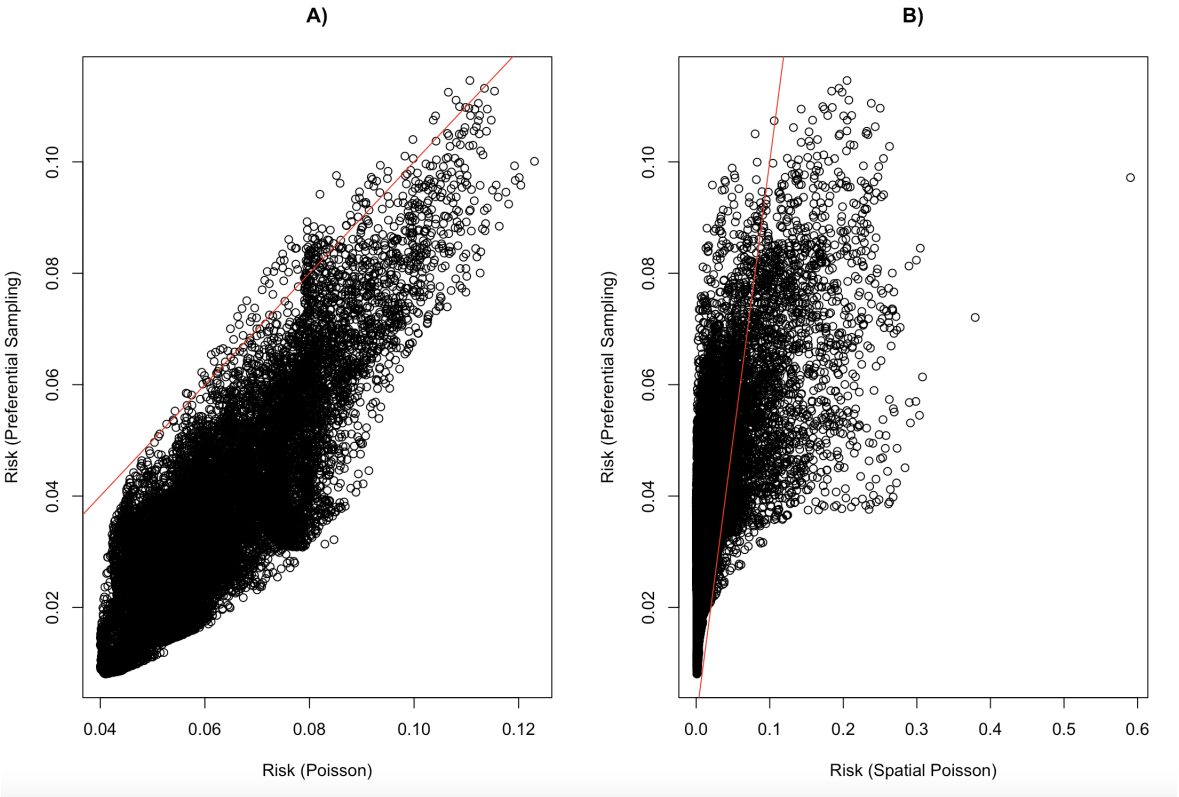

Figure 1.22: Comparison of the estimated plague risk over all discretization cells of the study region. Each point compares the predicted plague risk from the preferential sampling model (1.1) against that of the A) Poisson model (1.5) and B) spatial Poisson model (1.6) for a given grid cell. The red diagonal is of slope 1 with intercept 0.

To re-enforce the per-cell risk differences between model (1.1) and both reference

models, we observe that on average the mean absolute percent (MAP) difference in the per-cell risk values between models (1.1) and (1.5) is 47.375%, with a maximum MAP difference of 80.443%, while the MAP difference relative to model (1.6) is over 696%, with a maximum difference of over 36,000% (Table 1.7).

Model	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
(1.5)	0.021	36.008	49.502	47.375	59.923	80.443
(1.6)	0.007	140.777	396.454	696.135	755.482	36089.364

Table 1.7: Summary of mean absolute percent differences in predicted Sciurid plague risk by the preferential sampling model (1.1) in comparison to the Poisson (1.1) and spatial Poisson (1.6) models. The absolute percent difference is calculated on a per-cell basis, as $|(r_{i,2} - r_{i,1})|/r_{i,2}$, where $r_{i,1}$ denotes the risk predicted by one of the reference models for the ith grid cell and $r_{i,2}$ denotes that of model (1.1).

We emphasize that the previous comparisons do not convey any notion of error with respect to the true, unknown, risk surface. That is, we cannot say from this data alone that the reference models (1.5) or (1.6) over- or underestimate the true disease risk for particular cells, but rather, that they show quantitatively different predictions of risk compared to our proposed method.

In addition to predicted risk, parameter estimates also differ greatly among the three models (Table 1.8). The greatest differences in estimates are witnessed for the intercept parameters of cases and controls, $\beta_{0,+}$ and $\beta_{0,-}$, respectively. The slope parameters, $\beta_{1,m}$ and $\beta_{2,m}$, corresponding to the first two PRISM principal components, show modest differences for the Poisson model and proposed model, but are widely divergent for those in comparison to the spatial Poisson model. For instance, the spatial Poisson model estimates $\beta_{1,+}$ to be 1.956, while the preferential sampling model and Poisson model estimate it to be 0.651 and 0.668, respectively.

Parameter	Model	Estimate	Variance
$\beta_{0,+}$	Poisson	1.181	0.0012
$\beta_{0,+}$	Spatial Poisson	-2.852	0.26642
$\beta_{0,+}$	Preferential Sampling	-1.923	0.036
$\beta_{1,+}$	Poisson	0.668	0.00075
$\beta_{1,+}$	Spatial Poisson	1.956	0.12131
$\beta_{1,+}$	Preferential Sampling	0.651	0.022
$\beta_{2,+}$	Poisson	0.557	0.00089
$\beta_{2,+}$	Spatial Poisson	0.692	0.13216
$\beta_{2,+}$	Preferential Sampling	0.295	0.011
$\beta_{0,-}$	Poisson	3.998	7e-05
$\beta_{0,-}$	Spatial Poisson	2.311	0.00384
$\beta_{0,-}$	Preferential Sampling	1.378	0.027
$\beta_{1,-}$	Poisson	0.415	5e-05
$\beta_{1,-}$	Spatial Poisson	0.673	0.00352
$\beta_{1,-}$	Preferential Sampling	0.314	0.017
$\beta_{2,-}$	Poisson	0.477	5e-05
$\beta_{2,-}$	Spatial Poisson	0.315	0.00613
$\beta_{2,-}$	Preferential Sampling	0.179	0.008

Table 1.8: Comparison of parameter estimates for the CDPH Sciurid analysis. Poisson estimates are taken as maximum likelihood estimates from model (1.5). Preferential sampling and spatial Poisson estimates are posterior means obtained from post-burnin MCMC samples, while variances for these two models are posterior variances.

Lastly we provide estimates for parameters specific to the proposed model (Table 1.9), namely preferential sampling parameters α_+, α_- along with the range (θ) and marginal variance (ϕ) associated with the covariance function of the Gaussian process. The preferential sampling parameters are estimated as $\alpha_+ = 1.429$ and $\alpha_- = 1.297$. The importance of these parameters lies in the way in which they modulate the impact of the shared latent process in the predicted risk, or in other words, help quantify the impact of preferential sampling. We also estimate the spatial range range to be 1.800, with posterior variance 1.757, and marginal variance to be 16.987, with posterior variance 8.377.

Parameter	Posterior Mean	Posterior Median	Posterior Variance
α_+	1.429	1.435	0.003
α_-	1.297	1.298	0.002
Range	1.8	1.757	0.124
Marginal Variance	16.987	16.642	8.377

Table 1.9: Model (1.1) parameter estimates for the Sciurid plague analysis. Posterior means, medians and variances are calculated from post-burnin samples. The + and − subscripts denote disease positive and negative status, respectively.

1.5.4 Discussion

This analysis fit the proposed preferential sampling method and two benchmarks to the CDPH plague surveillance dataset, with the goal being to determine whether modeling the sampling process could impact the predicted risk map. In addition, it was also of interest to determine the degree to which the data are, in fact, preferentially sampled in the strict sense - that is, through stochastic dependence between the choice of sampling locations and risk for plague - beyond the general sense in which it is already known that the disease surveillance system tends to sample for plague in high risk regions. That is, if the tendency to sample for plague in high risk regions can be accounted for by fixed effect covariates alone, then the methodological innovations featured in this project would prove unnecessary, at least for this particular dataset. But on the contrary, the results of this analysis lend support toward the opposite conclusion, that additional measures are indeed necessary to account for preferential sampling.

Firstly, the evidence shows that the proposed preferential sampling model does indeed produce considerably different risk maps in comparison to the reference methods, i.e., Poisson regression and spatial Poisson regression. The per-cell risk scatterplots (Figure 1.22A) show that the Poisson model yields higher estimates of risk for most cells relative to model (1.1), but also predicts lower risk for cells calculated to be high risk by model (1.1). These discrepancies are readily apparent in Figure 1.20.

Most notably, the Poisson model tends to show a higher level of disease risk in the Central Valley region of California as well as in the northeastern portion of the map, likely due to the fact that the Poisson model estimated a much greater case intercept $\beta_{0,+}$ and lower control intercept $\beta_{0,-}$ than the preferential sampling model (Table 1.8). This elevated estimation of risk is consistent with what would be expected from modeling preferentially sampled data with traditional approaches. The fact that the data are sampled predominantly in high risk regions is expected to result in elevated predictions of risk at unsampled areas. There are also particular pockets along the coastline where model (1.5) predicts higher disease risk. But far from it being the case that model (1.5) predicts higher risk everywhere, the scatterplot of per-cell risk differences (Figure 1.22A) shows areas where the proposed model predicts higher risk than model (1.5). Further investigation of these cells showed their locations to be in 4 small pockets along the eastern Sierra Nevada mountains, and 3 pockets in southern California. These difference are noteworthy because they demonstrate that even preferential sampling can lead to potentially missing areas of the highest risk, when the data are analyzed by traditional methods. However we emphasize the adverb potentially, given that we do not know whether the higher risk regions as predicted by model (1.1) are truly higher risk, only we note that if model (1.1) is correct, then the Poisson model would miss these regions.

The proposed method shows even greater divergence in predicted risk when compared to the spatial Poisson model (1.6). Unlike the Poisson model, which demonstrated at the very least a general directional trend of agreement in per-cell risk with model (1.1) (1.22A), the spatial Poisson model calculated both greatly higher and lower values than model (1.1) (1.22B). Areas of exceptionally predicted high risk (> 0.5) are apparent in the risk map produced by model (1.6) (Figure 1.22B), which are completely absent from the map of model (1.1). In addition, the spatial Poisson map shows much higher risk in clusters along the Sierra Nevada mountain range, in the

eastern and central portion of the map, compared to model (1.1). Predicted risk values in excess of 0.5 are so high as to be considered unreasonable model outputs given the known low prevalence of plague in Sciurids. Hence, it is important to understand how and why these excessive values arise in model (1.6).

The general tendency of the spatial Poisson model is to predict small areas of sharply elevated risk, which quickly decline to values near zero. This phenomenon can be explained by considering the raw disease prevalences and case counts in the cells predicted to be of highest risk by the spatial Poisson model (Table 1.10). Here, extremely high raw prevalences (i.e. the number of per-cell cases divided by the total number of observed specimen), between 0.368 and 1, arise from low overall numbers of observed specimen, on the order of 1 to 14 total observed rodents. The resulting spatial random effects are of high positive value, for cases (at most 2.273), and of near zero or low negative value, for controls (between -0.256 and -3.432). If we let w_+ and w_- denote the random effects for cases and controls, respectively, then since since disease log odds are given by the difference of case and control log intensities, we can see that the elevated risk predictions are driven by the differences $w_+ - w_-$, which are relatively large for these cells.

Predicted Risk	Prevalence	Cases	w_+	w_-
0.713	1	2	0.12	-3.432
0.569	1	1	1.062	-3.033
0.511	1	1	-2.717	-4.22
0.489	0.429	3	2.493	-1.37
0.424	0.368	14	2.273	-0.256

Table 1.10: Predicted risks, raw disease prevalences, case counts and random effect values from the spatial Poisson model for the raster cells with the top 5 greatest predicted risks. w_+ and w_- denote the spatial random effects from model (1.6) for cases and controls, respectively.

In contrast, the proposed model operates under a completely different framework with regard to spatial random effects, the details of which explain why model (1.1)

does not exhibit the same excessive prediction of risk for these cells. We recall that model (1.1) does not assign case and control specific Gaussian processes to capture unexplained spatial variation, as does model (1.6). Rather, a single Gaussian process w is shared among the component of the model describing the distribution of sample sites and the components of the model describing abundances of cases and controls. The impact of w is brought about in the case and control intensity functions by means of the product $\alpha_+ \times w$, for cases, and $\alpha_- \times w$, for controls. As Table 1.11 shows, for the cells predicted to have the highest risk by model (1.6), in the framework of model (1.1), the differences $(\alpha_+ \times w) - (\alpha_- \times w)$ are in fact relatively minor, resulting in low predicted risks. Thus, due to the fact that w is shared between cases, controls, and locational information, there is no opportunity for random effects to inflate in positive and negative directions in response to high case counts and low control counts, from which it follows that the predicted risks are kept in check.

Predicted Risk	Prevalence	Cases	$\alpha_+ \times w$	$\alpha_- \times w$	$(\alpha_+ \times w) - (\alpha_- \times w)$
0.04	1	2	-3.65	-3.312	-0.338
0.034	1	1	-1.328	-1.205	-0.123
0.026	1	1	-1.511	-1.371	-0.14
0.031	0.429	3	-6.328	-5.742	-0.586
0.019	0.368	14	-5.827	-5.288	-0.539

Table 1.11: Predicted risks, raw disease prevalences, case counts and random effect values from the proposed model (1.1) for the raster cells from Table 1.10 which the spatial Poisson model (1.6) predicts to have the highest risk.

We now consider the parameter estimates of β_+ and β_- obtained from each model (Table 1.8). However, it is quite natural to call into question the meaning and usefulness of attempting to interpret any single value of β_+ or β_-, given that these parameters do not directly correspond to actual physical measurements, but rather to the first two principal components of the dimensionally reduced PRISM climatic variables. The slope associated with a principal component is arguably of little scientific interest per se. While that judgment may be so, an examination of these parameters is still

worthwhile for the purpose of explaining the differences in the predicted risk maps calculated by each model. Firstly, there are clear and substantial differences in the parameter estimates from each model. The greatest differences arise in the case and control intercepts, $\beta_{0,+}$ and $\beta_{0,-}$. The Poisson model shows notably greater intercepts than both other models, as would be expected for two primary reasons, firstly, due to the fact that if samples tend to be conducted in areas at high risk for plague, then an upward pressure is naturally exerted on the intercept of cases, and if these high risk areas tend to be where Sciurids are more likely to be found overall, then a similar pressure would act on the intercept of controls. The differences may also be attributed to model construction, insofar as models (1.1) and (1.6) both have additional components accounting for Sciurid abundances, namely $\alpha_+ \times w$ and $\alpha_- \times w$ for model (1.1), or w_+ and w_- for model (1.6). Intercepts are not the only parameters which show notable differences across models. The fact that all slope parameters estimated by model (1.6) are substantially different than those from the other models is perhaps related to the erratic behavior of the spatial random effects in this model, as previously observed. While the differences in slopes between the Poisson and preferential sampling models are much less pronounced, they are nonetheless nontrivial given the tight variances around these estimates, which may in part explain differences in the predicted risk maps (Figure 1.20). However, the biggest portion of the variability in risk maps between these two models is most likely explained by estimates of the preferential sampling effects, $\alpha_+ \times w$ and $\alpha_- \times w$, which we now discuss.

Estimates of α_+ and α_- are crucial for the interpretation of this analysis since these parameters modulate the effect of the spatial random process w on the observed abundances of cases and controls. The α parameters can be seen as representing the strength of preferential sampling present in the data through their relationship to w. For instance, values of $\alpha_+ = 0, \alpha_- = 0$ imply that the latent process w has no bearing on case or control abundances, and hence, there is no stochastic relationship between

the pattern of observation sites and disease risk, or in other words, that the data are not preferentially sampled. Alternately, if $\alpha_+ > 0, \alpha_- > 0$, and $\alpha_+ = \alpha_-$, then the sampling mechanism corresponds to a scenario wherein observations tend to be conducted in areas of higher case and control counts, but not in areas of higher disease risk. But if the data are preferentially sampled then the disease log odds at location x are calculated by model (1.1) as $z_\lambda(x)^T\beta_+ + \alpha_+ \times w - z_\lambda(x)^T\beta_- - \alpha_- \times w$, which implies that as w increases, in order for disease log odds to increase, $\alpha_+ \times w - \alpha_- \times w$ must increase, and consequently, that it must be the case that $\alpha_+ > \alpha_-$. The reverse line of reasoning holds to show that $\alpha_+ > \alpha_-$ implies that the data are preferentially sampled, with the degree of preferential sampling increasing as α_+ grows larger than α_-. Thus, preferential sampling is evidenced by whether $\alpha_+ > \alpha_-$, and the degree or strength of preferential sampling by the magnitudes of $\alpha_+ \times w - \alpha_- \times w$ relative to the other components of the disease log odds construction, i.e. $z_\lambda(x)^T\beta_+ - z_\lambda(x)^T\beta_-$.

Estimates, in the form of posterior means, for α_+ and α_- are 1.429 and 1.297, respectively (Table 1.9). Given the posterior distributions of each parameter, Bayesian hypothesis testing easily confirms that $\alpha_+ > \alpha_-$ at the 95% confidence level. We interpret this result as evidence for preferential sampling in the dataset. The question becomes how to quantify the strength of preferential sampling apparent in this surveillance system. A good quantification strategy is to calculate what portion of the linear predictor of log disease odds is comprised by the preferential sampling terms, $\alpha_+ \times w - \alpha_- \times w$. That is, for each grid cell location x in the study region, we calculate

$$100 \times \frac{|\alpha_+ \times w(x) - \alpha_- \times w(x)|}{|z_\lambda(x)^T\beta_+ - z_\lambda(x)^T\beta_-| + |\alpha_+ \times w(x) - \alpha_- \times w(x)|}$$

where we take absolute values $|.|$ in order to facilitate the calculation of summary statistics (Table 1.12). The expectation of this quantity over all grid cells can be

thought of as the expected percent of the linear predictor of log disease odds that is attributable to the spatial process w. We see that on average, roughly 8% of the linear predictor log disease odds originates from the preferential sampling component. While this percentage may seen minor, roughly 25% of cells have over 10% of the linear predictor comprised by the preferential sampling component. By this metric, and by the comparison of the per cell risk predictions, we conclude that there is a mild to moderate effect of preferential sampling on the data.

Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
0	4.00	8.00	8.71	13	23

Table 1.12: Summary of estimated percentage of log disease odds attributable to preferential sampling. Summaries are calculated over per cell log disease odds. 1st and 3rd Qu. denote the first and third quartiles.

This analysis has provided several key insights into the spatial characteristics of plague and the dynamics of the sampling mechanism behind the surveillance system, along with other points of scientific interest. Firstly, under every method considered, there is notable spatial heterogeneity in the risk of plague. Secondly, there are indeed significant relationships between plague and climate which, in part, explain this heterogeneity. But the most important conclusions center around the effect of preferential sampling on the predicted risk of plague. Most importantly, we can conclude that there is evidence of an impact of preferential sampling on predicted risk, as shown by the unique features of the predicted risk map generated by the proposed method. However, a key limitation of this evidence is that the comparisons of estimated risk between models do not yet account for the posterior variances associated with these estimates. One obvious concern is that while point estimates of risk may differ on the surface, such apparent differences may be due to statistical noise. We clearly need to impose statements of statistical significance around our comparisons. However, to do so is beyond the capability of the downscaling technique featured in this project. In particular, our current downscaling method generates merely point estimates of risk

at high resolution locations, but not arrays of posterior samples at these locations. Without posterior samples, statements about the statistical significance of differences in risk at these sites become impossible. A new downscaling approach is needed to address this issue, which we introduce in Project 3. But for this project, we present the initial, exploratory evidence that preferential sampling impacts the estimated risk of plague. More definitive evidence supporting this point is presented in Project 3.

Two additional limitations of this analysis are also apparent. Firstly, the data have been aggregated across all Sciurid species, which may mask heterogeneity in risk and sampling dynamics particular to certain species. Secondly, the data have been aggregated over a large span of time, from 1983 to 2015, which has the potential to miss temporal trends in the disease and sampling processes. We address these concerns in future chapters of this dissertation.

Chapter 2

A Multivariate Framework to Address Preferential Sampling

2.1 Introduction

A key motivation of multivariate geostatistical modeling is to exploit shared statistical information between different responses to make better predictions. For instance, suppose an observation system collects abundant measurements of one response over a broad spatial extent, but only sparsely gathers data on another type of response. If the two responses are well correlated, analyzing them as if they are independent would miss a valuable opportunity to borrow information from the well sampled response to inform predictions made about the sparsely observed response. In contrast, joint analysis would take advantage of this correlation to better predict values of both response types. In Project 2, we bring the benefits of the multivariate approach to bear on the focus of our previous project, namely preferential sampling in the surveillance of zoonotic diseases. Specifically, we develop a new multivariate geostatistical model

which corrects for preferential sampling when estimating the risk surfaces of a disease common across multiple species, and which shares information between species in a hierarchical modeling framework. In this project we conduct two extensive simulation studies, one comparing the predictive performance of our proposed model relative to a univariate approach, and another probing the robustness of our model to violations of a key assumption, namely separability, underpinning the construction of its multivariate structure. We conclude with an application to a disease surveillance dataset monitoring the occurrence of plague in Sciurids (the rodent family of squirrels) and coyotes across the state of California.

The key innovation of Project 2 consists of a multivariate approach to correct for preferential sampling. Preferential sampling refers to a scenario in which there is statistical dependency between the locations at which data are recorded and the values of response measured at those locations. This mode of sampling typically arises in applications where observation sites are assigned to areas which are of high value for the response of interest, such as in environmental health monitoring (Lee et al., 2011), where the goal may be to monitor pollutant levels in the most damaged areas, species distribution modeling (Gelfand and Shirota, 2019), in which it is most practical to search for a species in areas where it is most likely to be found, or disease surveillance (Cecconi et al., 2016), wherein limited resources are best assigned to areas at highest risk for a disease.

The core problem addressed in Project 1 was the liability of traditional geostatistical methods to yield biased statistical predictions of the response when fit to preferentially sampled data (Diggle et al., 2010; Lee et al., 2011; Gelfand et al., 2012; Lee et al., 2015). While previous research has proposed solutions to correct for preferentially sampling (Diggle et al., 2010; Pati et al., 2011; Lee et al., 2011; Lee et al., 2015), few methods have been developed to accommodate disease surveillance data. Moreover,

the exceptions to this trend have typically resided in veterinary health monitoring (Cecconi et al., 2015; Rinaldi et. al., 2014), and as such, have taken advantage of particular modeling simplifications which may not be available in other applications, such as cases of disease arising within a known, fixed set of points (i.e. farms). The primary contribution of Project 1 was thus to correct for preferential sampling in a more generalized disease surveillance setting. Project 2 extends the framework of Project 1 to encompass disease surveillance data from multiple species, in effect sharing information between species in a joint modeling framework.

Our basic strategy in this project is similar to that in Project 1, but generalized to multiple different species at risk for the disease of interest. As before, we model the distribution of observation sites along with case and control abundances in terms of a shared latent spatial process, only in this context, the process is multivariate, consisting of distinct but correlated processes for each species. In this way we can share information between species through the correlations among these processes, while still preserving enough flexibility to capture ecological and epidemiological idiosyncrasies specific to each species.

This analysis considers a disease surveillance application encompassing multiple species, with the intent to apply the joint model proposed in the methods section of Project 2. As in Project 1, the disease surveillance system of interest is operated by the California Department of Public Health (CDPH), targeting plague (infection with *Yersinia pestis*) among its animal hosts across the state of California. However, in this analysis we consider not only data pertaining to Sciurids, the rodent family of squirrels, but also to coyotes, both of are monitored by preferential sampling mechanisms. Taken together, the surveillance data from coyotes and Scurids provide a much greater spatial coverage of the study region than that from any one species alone (Figure 2.1), and thus may offer improve estimates of the disease risk surfaces

when analyzed together.

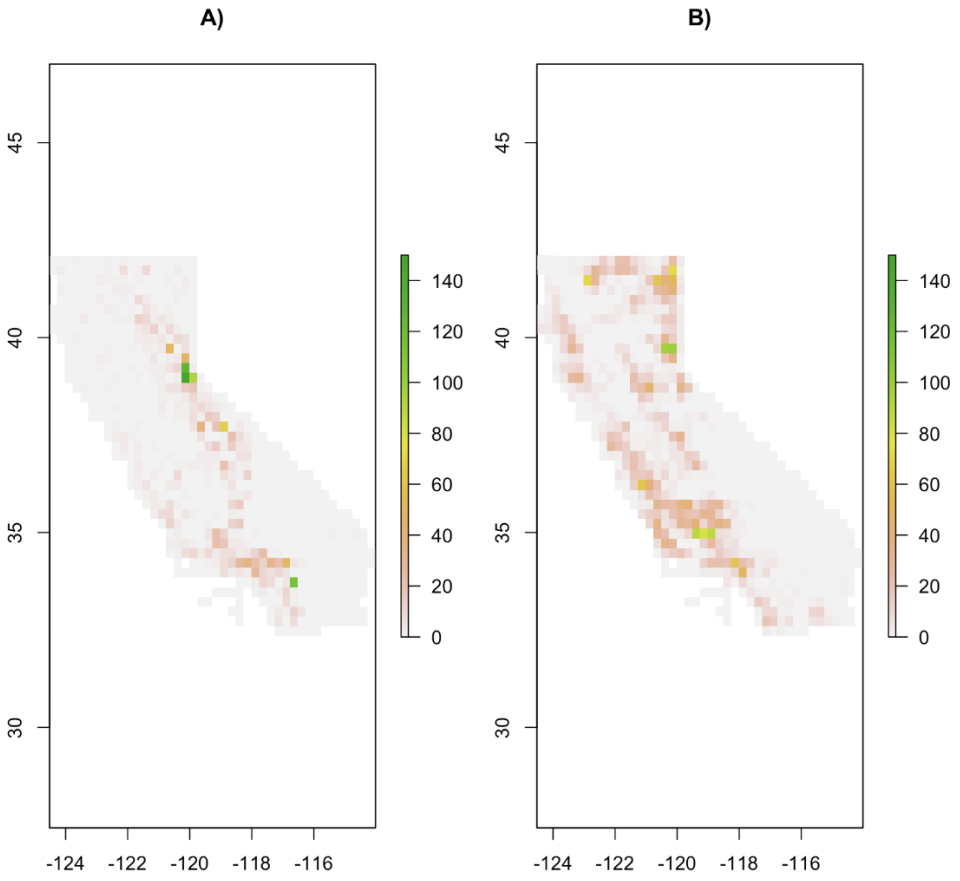

Figure 2.1: Distributions of distinct sampling locations for A) rodents and B) coyotes between 1982 and 2015.

The first species grouping monitored by the CDPH plague surveillance system is the Sciurid family, encompassing 21 different species in the observed data (i.e. Antelope Ground Squirrel, Antelope Ground Squirrel, Belding's Ground Squirrel, California Ground Squirrel, Chipmunk, Least Chipmunk, Long-eared Chipmunk, Lodgepole Chipmunk, Merriam's Chipmunk, Panamint Chipmunk, Shadow Chipmunk, Siskiyou Chipmunk, Sonoma Chipmunk, Uinta Chipmunk, Yellow-pine Chipmunk, Golden-mantled Ground Squirrel, Ground Squirrel, Yellow-bellied Marmot, Pine Squirrel, and Squirrel). The surveillance system collects data by conducting a series of sampling events at locations throughout California, in which Sciurids are trapped and

subsequently tested for *Yersinia pestis*. The data contain samples collected between 1983 and 2015. This analysis aggregates data for all Sciurid species, and for all years observed. The surveillance system predominantly collects data by a strategy of preferential sampling, assigning sampling locations to high risk or high impact areas. Here, risk is assessed to be high in what are viewed as plague endemic regions, as determined by historic cases of plague in humans or recovered Sciurid specimen, and high impact areas are regions where cases of plague in humans would be particularly likely and damaging, such as in national parks or areas which are climatically suitable for plague and have high human usage, such as the Lake Tahoe area in the north eastern portion of the state.

The sampling mechanism for coyotes (*Canis latrans*) follows a different chain of events than that of the Sciurids but one which nevertheless is an instance of preferential sampling. In the first step of the sampling process, coyotes are recovered either in the form of roadkill or in response to reported harassment of livestock. In either case, blood samples are collected and submitted to the CDPH along with locational identifiers describing the point of recovery. These identifiers vary in precision from high quality, such as latitude and longitude coordinates obtained from GPS, to verbal directions, such as the estimated mileage from a nearby road or other identifier. We note here the potential for misclassification error in the latter form of description, due to the fact that some verbal directions may be simply incorrect or notably different than the true point of recovery. Upon reception of a blood sample, CDPH conducts F1 antigen blood tests for plague only if the carcass was recovered in what is deemed to be a plague endemic region. This final step, the conduction of tests conditional on origination from a plague endemic area, is a form of preferential sampling, as tests for plague are thus conducted only in at risk areas.

The objective of our analysis is thus to estimate plague risk maps for Sciurids and

coyotes through a joint modeling framework which adjusts for preferential sampling. These last two points, joint modeling and preferential sampling, are key areas of focus in our discussion. Regarding the first, we wish to ascertain the strengths of correlations estimated between the two species groups, which are good indicators of how much information can be "shared" between groups, and thus a reflection of any gain to be acquired by using a joint model as opposed to a univariate approach. We emphasize that while we will ultimately produce separate risk maps for plague in coyotes and Sciurids, the map for each species grouping will in principle be open to statistical influence from information originating in the other species grouping. For the second point, we wish to characterize the degree or estimated effect of preferential sampling on the spatial predictions of risk.

Project 2 is organized as follows. In the following methods section, we present the relevant background on multivariate geostatistical methods and multispecies distribution models, before introducing our proposed method. We then report the results of 2 simulation studies comparing the performance of the new method to the benchmark of Project 1, and testing the robustness of the model to violations of the separability assumption. Lastly, we conclude with an analysis of the coyote and Sciurid data.

2.2 Methods

2.2.1 Introduction

Project 2 extends the previously developed preferential sampling methods of Project 1 to integrate multiple data types in a joint modeling framework. These new developments are motivated by the same plague surveillance application as in Project 1, but with the key distinction of creating a hierarchical model to ingest data from mul-

tiple disease host species. Our simulations consider two host species, corresponding to rodents and coyotes from the actual dataset, but in principle the modeling framework developed here can be extended to incorporate any number of host species. The essential novelty of our approach lies in the creation of this flexible joint modeling framework which corrects for preferential sampling and integrates disease surveillance data pertaining to multiple species.

2.2.2 Multi-species Distribution Modeling

The literature contains a substantial precedent of enhanced modeling performance obtained from multi-species analysis. In particular, we now turn our attention to the domain of species distribution modeling. Species distribution models (SDMs) seek to characterize the abundance and distribution of animal species over geographic areas (Elith and Leathwick, 2009). SDMs are fitted to data which typically possess many of the same complexities and challenges present in disease surveillance applications, especially when surveillance targets the presence of a disease in the animal population, such as the Plague surveillance system examined in the analysis chapter of Project 1, or when systematic surveys are prohibitively expensive, time consuming or logistically challenging (Rocchini et al., 2017). Convenience or opportunistic sampling (Fithian et al., 2015), preferential sampling (Michalcova et al., 2011; Pennino et al., 2019), detection error (Dorazio 2012), observational error (Royle et al., 2007; Cressie et al., 2009), uncertain absences or false absences (Royle and Link, 2006), and locational uncertainty (Mitchell et al., 2017) are all common issues complicating the analysis of species occurrence data.

Of interest to the objectives in Project 2 are those species distribution models which integrate multiple types of data in a joint modeling framework. A common theme of

data integration in SDMs is the challenge of combining what are known as *presence-only* and *presence-absence* data (Gelfand and Shirota, 2019). Presence-only data record only the locations where organisms from some species are observed, but not those locations at which surveys were conducted but no organisms were present (Renner et al., 2015). The pitfall of analyses solely reliant on such data is the fact that, given presences alone, key parameters in species distribution models relating to species abundance are not estimable (Fithian et al., 2015). In contrast, presence-absence data record both presences and absences of a species across some set of survey sites, typically distributed throughout the study region according to a pre-defined survey design. Presence-absence data solve the problem of estimability which arises in presence-only analyses. However, presence-absence surveys typically lack the large spatial coverage which characterizes presence-only datasets. A solution combining the strengths of both data types is to formulate a joint model encompassing presence-only and presence-absence data (Fithian et al., 2015; Gelfand and Shirota, 2019). We now review several of the key models designed for this purpose.

In a classic example resembling our own application to Plague surveillance, Fithian et al. (2015) pooled presence-only and presence-absence data from multiple Eucalyptus tree species in a joint model, finding that the joint analysis offered a superior predicted species distribution surface than models fit to individual species alone. The basic scenario addressed consists of a combination of presence-only reportings of species locations, along with presence-absence surveys conducted for some subset of species. Complicating the analysis is the fact that the presence-only data are recorded through a biased collection mechanism known as *opportunistic sampling*: wherein samples are more likely to be collected particular areas of the study region, often as a result of accessibility or convenience. Consequently, the authors model the presence-only data as a thinned point process. For each species k, the true underlying point process describing the species' locations is assumed to be an inhomogeneous Poisson process,

$S_k \sim \text{IPP}(\lambda_k)$, and the observed point process of locations is a thinned inhomogeneous Poisson process $T_k \sim \text{IPP}(\lambda_k(s)b_k(s))$. Intensity functions are assumed log-linear in spatially varying coefficients $x(s)$ and $z(s)$, summarized as:

$$
\begin{aligned}
S_k &\sim \text{IPP}(\lambda_k) \\
T_k &\sim \text{IPP}(\lambda_k(s)b_k(s)) \\
\log(\lambda_k(s)) &= \alpha_k + x(s)^T\beta_k \\
\log(b_k(s)) &= \gamma_k + z(s)^T\delta
\end{aligned}
$$

Since δ is not indexed by k, it allows for the pooling of information across multiple species. Rewriting the log-intensity function of the observed point process as $\lambda_k(s)b_k(s) = \alpha_k + x(s)^T\beta_k + \gamma_k + z(s)^T\delta$, we encounter a problem. From presence-only data alone the sum $\alpha_k + \gamma_k$ is identifiable, but not the individual terms α_k and γ_k. However, presence-absence data may come in the form of abundances of species k in survey plot i, denoted N_{ik}, or occupancy measures, $y_{ik} \in \{0, 1\}$. For survey abundance data around small plots A_i with center points s_i,

$$
\begin{aligned}
N_{ik} &\sim \text{Poisson}(|A_i|\lambda_k(s_i)) \\
&= \text{Poisson}(|A_i|\exp(\alpha_k + x(s_i)^T\beta_k))
\end{aligned}
$$

Crucial here to addressing sampling bias is the fact that the distribution of N_{ik} does not depend on the bias process, thus resolving the lack of identifiability of α_k and γ_k when maximizing the joint log-likelihood of the presence-only and presence absence

data, written as

$$\ell(\alpha, \beta, \gamma, \delta) = \Sigma_k \ell_{k,\text{PA}}(\alpha_k, \beta_k) + \Sigma_k \ell_{k,\text{PO}}(\alpha_k, \beta_k, \gamma_k, \delta)$$

where $\ell_{k,\text{PA}}$ and $\ell_{k,\text{PO}}$ are the log-likelihoods of the presence-absence and presence-only data for the kth species, respectively, which may be combined additively in the joint log-likelihood due to the fact that the presence-only and presence-absence data are collected independently, across all species. Through adjusting for sampling bias by pooling presence-only and presence-absence data, the joint model substantially improved the out of sample predictive performance for 36 eucalypt species in south-eastern Australia compared to other methods, namely analyzing pooled presence only/presence-absence data for a single species, and pseudo-absence regression of the presence-only data alone. Moreover, the bias adjustment worked even for a species that had no-presence absence data by borrowing information from other species.

Similar examples abound in the literature. Dorazio (2014) developed a hierarchical point process model which also pooled data from opportunistic and planned surveys, but one that not only corrected for sampling bias but also for *detection error*, that is, failure of the observation process to record true presences of the species of interest. Hooten et al. (2011) combined multiple surveys of forest disease incidence in a hierarchical model to adjust for sampling bias. Chakraborty et al. (2011) included (known) sampling effort to offset the intensity function of observed data. Several other multi-species data integration solutions for reducing bias have been proposed (Phillips et al., 2009; Fletcher et al., 2016; Giraud et al., 2016), which are thoroughly reviewed in Hefley and Hooten (2016). Project 2 contributes to this body of work by integrating multi-species, preferentially sampled disease surveillance data.

2.2.3 Multivariate Gaussian Processes

A key motivation of multivariate geostatistical modeling is to exploit shared statistical information between different responses to make better predictions. We now consider multivariate extensions of the spatial process model introduced in Project 1:

$$Y(s) = \mu(s) + w(s) + \epsilon(s)$$

which describes a univariate response $Y(s)$ at location $s \in \mathbb{R}^2$ in terms of a mean component, $\mu(s)$, typically linear in fixed covariates $x(s)\beta^T$, along with a smooth spatial process $w(s)$ and optionally a spatially unstructured residual term, $\epsilon(s)$, referred to as the "nugget" effect. In many spatial applications $w(s)$ is taken to be a Gaussian process with known covariance function $k(s, s')$.

In the multivariate case, $Y(s)$ is now a $p \times 1$ vector of random variables observed at location s. Supposing that the process $Y(s)$ is observed at n different locations, and letting $Y = (Y(s_1), \ldots, Y(s_n))$, then the cross-covariance matrix, defined as $C(s, s^*) = \text{cov}(Y(s), Y(s^*))$, gives the covariance matrix between the response vector at site s and that from some other site s^*. It is required that for any arbitrary number n of and choice of locations, the resulting $np \times np$ covariance matrix for Y, denoted Σ_Y, must be positive definite. In the following three sections we detail alternate strategies for ensuring the positive definiteness of Y, namely separable models, linear models of coregionalization, and convolution methods. Model descriptions here are primarily taken from the extensive summary provided by Banerjee, Carlin and Gelfand (2004), unless otherwise referenced.

Separable Models

If ρ is a valid correlation function for a univariate spatial process, then one valid cross-covariance function takes the form

$$C(s, s') = \rho(s, s') \cdot T$$

where T is a $p \times p$ matrix interpreted as the covariance matrix of $Y(s)$ at a single point in space, and ρ acts to decrease the association between $Y(s)$ and $Y(s')$ as their distance increases (Banerjee, Carlin and Gelfand, 2004). In this case the covariance matrix of Y is $\Sigma_Y = H \otimes T$, where $H_{ij} = \rho(s_i, s_j)$ and $\otimes$ is the Kronecker product of two matrices. From this expression we know Σ_Y must be positive definite, given the fact that the Kronecker product of two positive definite matrices is positive definite, and H as well as T are both positive definite. This expression for Σ_Y makes computation more efficient due to the fact that $\Sigma_Y^{-1} = H^{-1} \otimes T^{-1}$ and $|\Sigma_Y| = |H|^p |T|^n$, which thus allows the calculation of inverses and determinants of $p \times p$ and $n \times n$ matrices, rather than an $n \times p$ matrix.

Estimation of the T matrix can be carried out with Gibbs sampling, provided T is assigned an inverse-Wishart prior distribution (Banerjee, Gelfand, and Polasek, 2000). It is noteworthy however that this result would not hold if a nugget effect, $\epsilon(s)$, of independent and identically distributed residuals were included in the spatial process model. The derivation of the conditional distribution of T given spatial random effects w, provided by Banerjee, Gelfand and Polasek (2000), is as follows.

Suppose $Y(s) = \mu(s) + w(s)$ is the multivariate response of p measurements observed at location s, that measurements are conducted at a total of n observation sites, and that $U \sim N(\mu, \Sigma_U)$, where $\mu = (\mu(s_1), \ldots, \mu(s_n))$ and $\Sigma_U = H \otimes T$. That is,

$$f(U|\phi,T) \propto |H \otimes T|^{-1/2}\exp(-0.5(U-\mu^T(H\otimes T)^{-1}(U-\mu)))$$

.

Furthermore, suppose that T has an inverse-Wishart Distribution $IW_k(r,\Omega)$ given by

$$f(T) \propto |T|^{-(r+p+1)/2}\exp(-0.5 \times \mathrm{tr}(\Omega T^{-1}))$$

Consequently

$$\begin{aligned}
&f(T|U,\phi) \propto |T|^{-(r+p+1+n)/2}\exp(-0.5[(U-\mu)^T H^{-1}\otimes T^{-1}(U-\mu)+\mathrm{tr}\Omega T^{-1})]) \\
&= |T|^{-(r+p+1+n)/2}\exp(-0.5[\Sigma_i\Sigma_j(H^{-1})_{ij}(U(s_i)-\mu(s_i))^T T^{-1}(U(s_j)-\mu(s_j))+\mathrm{tr}\Omega T^{-1}]) \\
&= |T|^{-(r+p+1+n)/2}\exp(-0.5[\Sigma_i\Sigma_j(H^{-1})_{ij}\mathrm{tr}(U(s_j)-\mu(s_j))(U(s_i)-\mu(s_i))^T T^{-1}+\mathrm{tr}\Omega T^{-1}])
\end{aligned}$$

Therefore $T|U,\phi \sim \mathrm{IW}(r+n, \Sigma_i\Sigma_j(H_{ij}^{-1}(U(s_j)-\mu(s_j))(U(s_i)-\mu(s_i))^T+\Omega)$. That is, the conditional distribution of T given U is also an inverse-Wishart distribution, with $r+n$ degrees of freedom and scale matrix $\Sigma_i\Sigma_j(H_{ij}^{-1}(U(s_j)-\mu(s_j))(U(s_i)-\mu(s_i))^T+\Omega$. Making use of the conditional distribution offers considerable gains in efficiency, as it is much faster to draw a $p\times p$ matrix from this distribution than to perform updates directly from the $np\times np$ covariance matrix Σ_Y.

While the separable model offers computational efficiency, it suffers from a number of limitations which limit the complexity of the spatial relationships it may capture. Firstly, since the cross-covariance function is composed of only one correlation function ρ, each latent dimension must have the same spatial range. In many practical

cases this may not be a reasonable assumption. Additionally, separable models restrict the correlation between measurements to tend to 1 as distance decreases, which may not be appropriate in cases where micro-scale variability persists. A nugget may be introduced to address this concern, but, as previously observed, doing so would prevent the Gibbs sampling of T. Lastly, the usual limitations associated with the assumptions of isotropy and stationary may come into play. For instance, in ecological applications it may be unreasonable to assume that the correlation of species abundances is the same at all regions of the study area, since 2 species may be closely linked in certain areas but not others.

Coregionalization

Linear models of coregionalization (LMC) offer a popular (Gelfand et al., 2004; Goulard et al., 1992; Journel et al., 1978; Grzebyk and Wackernagel, 1994; Wackernagel 1998) and more flexible approach to analyzing multivariate spatial responses, but at the expense of greater difficulty in model fitting. The underlying idea of coregionalization is to obtain dependent multivariate processes through transformation of independent processes. In the intrinsic LMC model introduced by Matheron (1982), the $p \times 1$ measured response $Y(s)$ at a particular location is formulated as the product of a full rank matrix A and a vector of spatial processes $w(s)$:

$$Y(s) = Aw(s)$$

where, in particular, the components of $w(s)$ are independent and identically distributed spatial processes. If these spatial processes are stationary, of mean zero and variance 1 with correlation function ρ, then $E[Y(s)] = 0$ and the cross-covariance matrix is

$$C(s, s') = \rho(s - s')AA^T$$

.

Letting $AA^T = T$, it becomes apparent that under these conditions the intrinsic LMC model is equivalent to the separable model described in the previous section.

A more flexible LMC is given by

$$Y(s) = Aw(s)$$

where now, the components of $w(s)$ are independent but not identically distributed spatial processes. A benefit of this formulation is that it enables each component of $Y(s)$ to have its own spatial range, an advantage not permitted by the separable model. If we specify $w_j(s)$, jth element of $w(s)$, to have mean μ_j, variance 1, and correlation function ρ_j, then $E[Y(s)] = A\mu$ for $\mu = (\mu_1, \ldots, \mu_p)$ and the cross-covariance matrix is

$$C(s, s') = \Sigma_{j=1}^{p} \rho_j(s - s')a_j a_j^T = \Sigma_{j=1}^{p} \rho_j(s - s')T_j$$

where a_j is the jth column of A, $T_j = a_j a_j^T$, and we denote $\Sigma_j T_j = T$. In practice, the multivariate spatial process $Aw(s)$ is typically incorporated into a more general modeling framework of the form

$$Y(s) = \mu(s) + v(s) + \epsilon(s)$$

where $Y(s)$ is the $p \times 1$ vector of observed processes, $\mu(s) = (x_1^T(s)\beta_1, \ldots, x_p^T(s))$ is a vector of spatially varying fixed effect covariates, $v(s) = Aw(s)$, and $\epsilon(s)$ is the white noise "nuggett" effect distributed as $\epsilon(s) \sim N(0, D)$, where D is a diagonal matrix with elements $D_{jj} = \tau_j^2$. For $Y(s)$ measured at an arbitrary n locations, the distribution of $Y = (Y(s_1), \ldots, Y(s_n))$ given fixed effects $\beta = (\beta_1, \ldots, \beta_p)$, white noise variance D, correlation functions $\rho = (\rho_1, \ldots, \rho_p)$, and matrix $T = \Sigma_j a_j a_j^T$ is

$$p(Y|\beta, D, \rho, T) \sim N\Big(\mu, \Sigma_{j=1}^p (H_j \otimes T_j) + I_{n\times n} \otimes D\Big)$$

for $\mu = (\mu(s_1), \ldots, \mu(s_n))$. While the general LMC offers greater modeling flexibility in comparison to the separable model, fitting the general LMC in a Bayesian framework has been remarked upon in the literature as a difficult and error prone process (Banerjee, Carlin and Gelfand, 2004). Typically, inverse gamma distributions are assigned to the diagonal elements of D, τ_j^2, which result in inverse gamma conditional distributions, thereby permitting D to be updated via Gibbs sampling. However, full conditional distributions for T and the range parameters of ρ do not fit into in standard form from any prior specification. Consequently, one typically resorts to the Metropolis-Hastings algorithm to estimate the parameters of ρ and T, a process far removed from the ease of fitting the separable model.

The final version of the LMC which we will remark upon is the spatially varying coregionalization model, written as

$$Y(s) = A(s)w(s)$$

where $A(s)$ is now a $p \times p$ matrix of spatially variable values, such as parametric functions of location. In this case, the cross-covariance function becomes

$$C(s, s') = \Sigma \rho_j(s - s') a_j(s) a_j(s')$$

which we see is dependent upon location, not merely the separation $s - s'$, and hence non-stationary. Thus, the spatially varying coregionalization model can be useful to capture complex spatial relationships, but with additional model fitting obstacles.

Convolution Methods

The last approach we consider relies on the convolutional representation of Gaussian processes to yield multiple, dependent processes, and has been widely used to describe multivariate geostatistical responses (Ver Hoef and Barry, 1998; Higdon 1998; Williams and Rasmussen, 1996). We recall that in the univariate setting, a Gaussian process can be obtained from the convolution of a white noise process with a kernel, which acts as a smoothing function (Higdon, 2002). That is,

$$v(s) = \int_S k(s - \tau) x(\tau) d\tau$$

for white noise process $x(s)$ and kernel k is necessarily a Gaussian process over some region S (Williams and Rasmussen, 1996). Convolution methods can give rise to Gaussian processes with correlated outputs. For instance, one multivariate convolution approach (e.g., Ver Hoef and Barry, 1998; Boyle and Frean, 2005) repeatedly convolves a spatial process to yield multiple, dependent spatial processes. In this methodology, $Y(s)$ is a p dimensional multivariate response measured at location s, and to each component $Y_j(s)$ kernel k_j is assigned. If the jth component of $Y(s)$ is defined as $Y_j(s) = \int_{R^2} k_j(u) Z(s + u) du$ for $k = 1, \ldots, p$ and for spatial process $Z(s)$ with correlation function ρ, then the resulting cross-covariance matrix for $Y(s)$ has

elements

$$C_{k,k'} = \sigma^2 \int_{R^2} \int_{R^2} k_j(s - s' + u) k_{j'}(u') \rho(u - u') du du'$$

which is a valid cross-covariance matrix. Typically the above integral cannot be analytically evaluated, hence the necessity to utilize numerical approximations. One benefit of this approach is that each component of $Y(s)$ is assigned its own kernel, allowing spatial range to vary among components. Offering even greater flexibility is the fact that since the convolution kernel may vary over space and time, Gaussian processes with non-stationary covariances may be obtained, as in Higdon (1998).

Several other approaches have been put forward to address dependence in the outputs of multiple spatial processes. One method is to pass independent Gaussian process priors through a softmax function (Williams et al., 1998), probit function (Kim et al., 2006; Girolami et al., 2006), or multiprobit function for ordinal regression (Chu et al., 2005). Another is found in Seeger et al. (2005), who propose a semi-parametric latent factor model with linear mixing for multivariate regression and classification problems.

Thus, we have reviewed a number of methods to address geostatistical processes with multiple, correlated outputs, increasing in complexity from the separable model, to the general linear model of coregionalization, and ultimately to the various convolution models presented above. For the stated objective of Project 2, forming a multivariate spatial model to correct for preferential sampling in the calculated risk maps of multiple species, we first adopt the separable multivariate Gaussian process out of consideration for model simplicity and comparative ease of fitting. We argue that the wisest approach for developing a new and experimental model such as this is to begin with the simplest level of complexity, assess where the model breaks, and

then introduce further extensions in response.

2.2.4 Proposed Method

The objective of Project 2 is to utilize disease surveillance data from multiple species in a hierarchical modeling framework, which shares information between species in the hope of offering improved risk maps, compared to what would be obtained by either estimating risk maps for each species via separate applications of model (1.1), or by ignoring inter-species differences and analyzing pooled data from all species with model (1.1).

Our basic strategy here is similar to that in Project 1, but generalized to multiple different species at risk for the disease of interest. As before, we model the distribution of observation sites along with case and control abundances in terms of a shared latent spatial process, only in this context, the process is multivariate, consisting of separate but correlated processes for each species. Specifically, we replace the univariate Gaussian process of model (1.1), $w(x)$, with a multivariate Gaussian process w, consisting of a separable covariance matrix intended to capture correlation in the underlying risk and sampling processes of each species. In this way we can share information between species through the correlations among these processes, while still preserving enough flexibility to capture species-specific variation. We opt here for the separable model due to the relative ease of updating the parameters of its covariance matrix, $H \otimes T$, in particular, relying on the convenience of Gibbs sampling the T matrix, thus avoiding the difficulties of fitting coregionalization models which have been remarked upon in the literature (Banerjee, Carlin and Gelfand, 2004).

We begin by discretizing the study region into a grid of equally sized cells, forming species-specific indicator variables denoting whether each cell contains an observation

event for each species. For example, Figure (2.2) shows a discretization of the state of California in which observed cells have been colored green, representing positive values of the indicator variables, and the true locations of observation sites, marked by red circles. We then sum case and control counts for each species across these grid cells, placing probability distributions on the values of the location indicators and case-control counts.

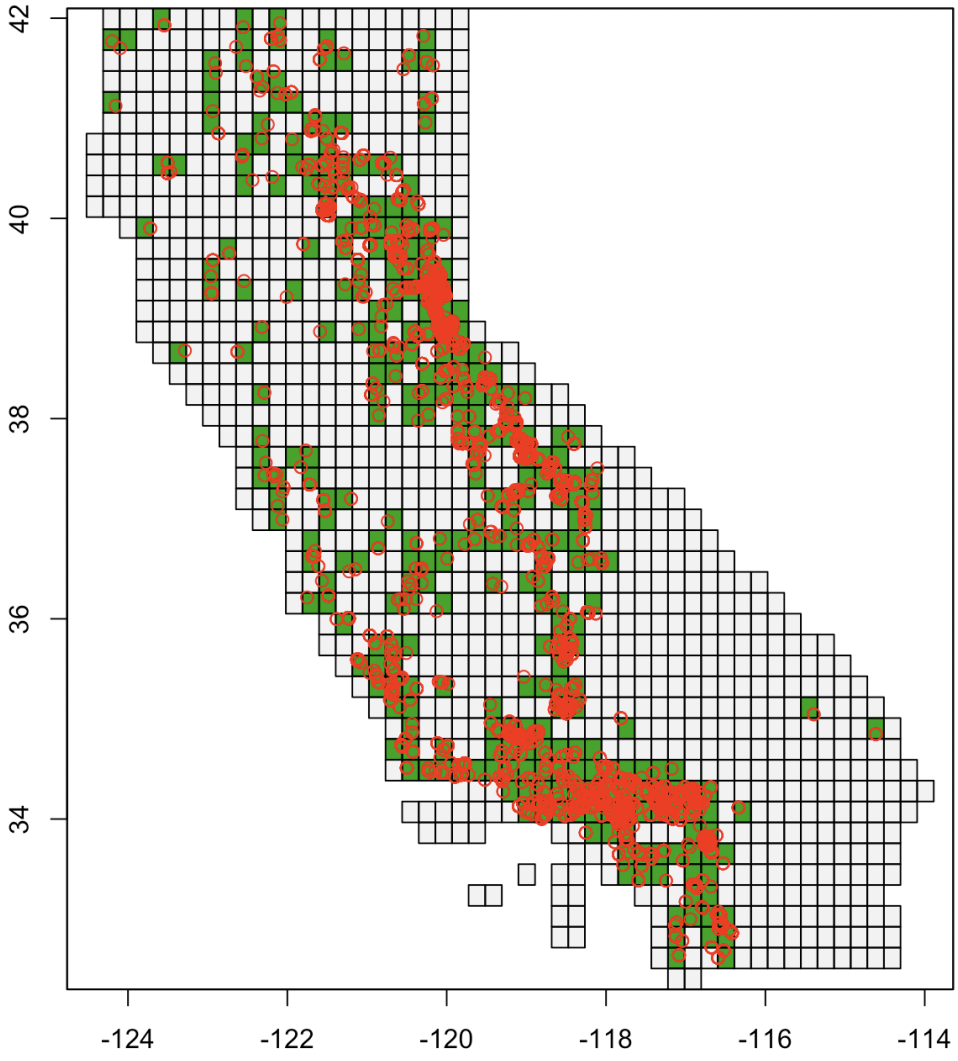

Figure 2.2: Discretized study region with observation status indicators, in green, identifying whether each cell contains an observation site for a particular species. True locations of observation sites are shown by red circles.

Formally, letting i index the grid cells in the discretization of the study region, k index the observed grid cells, m identify disease mark or status ($+$ for cases, $-$ for controls), $s \in \{1, \ldots, S\}$ index species, Y_{kms} denote the count of disease status m in kth observed grid cell for species s, and $\kappa_{is} \in \{0, 1\}$ identify whether the ith grid cell gets observed for species s, we propose the following model:

$$\begin{aligned}
&\kappa_{is}|\xi_s(x_i) \sim \text{Bernoulli}(\xi_s(x_i)) \\
&\text{logit}(\xi_s(x)) = w_s(x) \\
&w = (w_1, \ldots, w_s)^T, w \sim \text{MVGP}(0, H \otimes T) \\
&H_{ij} = k(x_i, x_j; \theta) \\
&Y_{kms}|w_s(x_i) \sim \text{Poisson}(\lambda_{ms}(x_k)) \\
&\log(\lambda_{ms}(x)) = z_\lambda(x)^T \beta_{ms} + \alpha_{ms} \times w_s(x)
\end{aligned} \tag{2.1}$$

Here, w is a multivariate Gaussian process, allowing each species to receive its own sampling process w_s while still capturing inter-species correlation via the separable covariance matrix $H \otimes T$. In this formulation, H is the spatial correlation matrix formed by correlation function k, with ijth element calculated as $H_{ij} = k(x_i, x_j; \theta)$. For our simulations and analyses we use the exponential correlation function

$$k(x, x'; \theta) = \exp(-||x - x'||/\theta)$$

and use the Euclidean distance $||x - x'||$ to measure the separation between points x and x'. This function is is stationary, i.e. its covariance depends only on the separation vector between two points, rather than their absolute location in space,

and isotropic, i.e. the covariance from point A to point B is the same as that from point B to point A. T is the $S \times S$ matrix capturing covariance between w_s values of different species in a single site. For greater flexibility we allow climatic covariate parameters β_{ms} to vary by species, as well as preferential sampling parameters α_{ms}.

A number of parallels between the multivariate model (2.1) and the single species model (1.1) of Project 1 can be drawn. Both models adopt the strategy of using a shared latent process, w, to correct for preferential sampling. The process is shared in the sense that it arises in both the locational and disease related components of the models, where the locational piece describes the distribution of observation sites and the disease related piece captures the abundances of observed cases and controls. A key distinction of the multivariate model is that is ascribes different, if dependent, processes to the pattern of observation sites for each species, namely κ_{is}, in contrast to model (1.1) which considers only a single set of observation indicators κ_i to encode the presence or absence of all observation sites in the dataset. The construction of the case and control log rate functions in both models is also very similar, given by $\lambda_m(x) = z_\lambda(x)^T\beta_m + \alpha_m \times w(x)$ in the single species model, and $z_\lambda(x)^T\beta_{ms} + \alpha_{ms} \times w_s(x)$ in the multivariate model. In the latter case, we allow fixed effect covariates β_{ms} and preferential sampling parameters α_{ms} to vary by species, along with the species specific random effects w_s. The ultimate goal of the introduced method is to thus correct for preferential sampling in calculating the risk maps for each species, as well as exploit inter-species correlations that may be present in the data to achieve better risk maps than analyzing species separately could provide.

To calculate the separate risk maps from model (2.1) we follow a similar line of reasoning as that used in Project 1, only in this case extended to yield different risk maps for each species of interest. Letting $r_s(x)$ be the probability an individual of species s at location x will be disease positive, or in other words, $r_s(x)$ be the risk

of the disease for species s at location $x \in \mathbb{R}^2$, the intensity functions of cases and controls for species s are, respectively

$$\lambda_{s,+}(x) = r_s(x)\rho_s(x)$$

$$\lambda_{s,-}(x) = c(1 - r_s(x))\rho_s(x)$$

where $\rho_s(x)$ is the population density function of species s and c is a constant determined by the study region. Consequently, the disease odds for species s, $r_s(x)/(1 - r_s(x)$ can be calculated up to a constant as

$$c^{-1}\frac{r_s(x)}{1 - r_s(x)} = \lambda_{s,+}(x)/\lambda_{s,-}(x)$$

Thus, solving the above for $r_s(x)$ and plugging in estimates of $\lambda_{s,+}(x)$ and $\lambda_{s,-}(x)$ yields an estimate the risk surface at point x. We repeat this calculation for each species $s = 1, \ldots, S$ in the dataset to yield a total of S separate risk maps. We emphasize here that, although the species of interest all must be susceptible to the same disease, the ultimate output of our analysis is not a single risk map but rather a set of risk maps specific to each species in the dataset.

2.2.5 Model Fitting

Parameters of the multivariate preferential sampling model (2.1) are estimated in a Bayesian approach encompassing a number of different sampling strategies, namely Hamiltonian Monte Carlo, Metropolis-Hastings random walk, and Gibbs sampling. Hamiltonian Monte Carlo is a Markov chain Monte Carlo technique that updates its state by simulating the dynamics of Hamiltonian particle physics. More specifically,

updates are influenced by the derivative of the log likelihood of the target distribution, which avoids the random walk behavior of the Metropolis-Hastings random walk algorithm, thereby more efficiently exploring the parameter space and also reducing correlation between samples. A complete technical description of Hamiltonian Monte Carlo is provided in the methods section of Project 1.

Here, in fitting model (2.1), independent normal prior distributions were assigned to β_{ms}, α_{ms}, with mean zero and relatively high prior variances. Hamiltonian Monte Carlo samplers are devoted to each of β_{ms}, α_{ms}, and w_s, where $s = 1, \ldots, S$ indexes species identity and $m \in +, -$ denotes case (+) or control (-) disease status. For instance, in the case of $S = 2$ species, a total of 10 different Hamiltonian Monte Carlo samples would be at play, 4 for the β_{ms}, 4 for the α_{ms}, and 2 for the w_s. The two tuning parameters for each sampler, step size ϵ and simulation length parameter L, representing the number of iterations to simulate Hamiltonian dynamics for each update step, were dealt with in separate manners. The step size was self-tuned using the dual averaging scheme developed by Hoffman and Gelman (2014), a technique that updates ϵ after each sample using a convex optimization algorithm to achieve a target acceptance rate, here set to 0.65 as recommended. The length parameter L was manually tuned, typically taking values between 8 and 10.

The spatial range parameter θ of the exponential correlation function, used to calculate the H matrix in the covariance matrix of the spatial random effects w in model (2.1), was estimated via Metropolis-Hastings random walk. We note that, due to the constraint of $\theta > 0$, the proposal distribution used to generate a proposed next value for θ was specified as the log-normal distribution, which has density function

$$q(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(- \frac{(\log(x) - \mu)^2}{2\sigma^2} \Big)$$

for $x > 0$. The mean of the proposal distribution was taken to be the log of the current value of θ. That is, given the gth MCMC sample $\theta^{(g)}$, the proposed next value in the Markov chain was drawn from

$$\theta^{(g+1)} \sim \text{Log-Normal}(\log(\theta^{(g)}), \sigma^2)$$

The proposal standard deviation σ^2 was manually tuned to yield acceptance rates close to 0.5. Given the asymmetric nature of the proposal distribution, the acceptance probability was calculated as

$$\min\Big(1, \frac{\ell(w^{(g)}; \theta^{(g+1)})}{\ell(w^{(g)}; \theta^{(g)})} \times \frac{p(\theta^{(g+1)})}{p(\theta^{(g)})} \times \frac{q(\theta^{(g)})}{q(\theta^{(g+1)})}\Big)$$

where $\ell(w^{(g)}; \dots)$ is the log likelihood of the spatial random effects w at the gth MCMC iteration, and $p(\dots)$ is the prior density of θ, taken here to be the gamma distribution, and $\frac{q(\theta^{(g)})}{q(\theta^{(g+1)})}$ is the ratio of log-normal densities from the current and proposed values of θ, which would have cancelled out had the proposal distribution been symmetric.

The cross-correlation matrix T of the spatial random effects w was assigned an inverse-Wishart prior distribution, $T \sim \text{Inverse-Wishart}(r, \Omega)$, which yields a conjugate posterior distribution given the random effect values. Consequently, T was updated via Gibbs sampling, drawing from the distribution

$$T|w \sim \text{IW}\Big(r + n, \Sigma_i \Sigma_j H_{ij}^{-1} w(s_j) w(s_i)^T + \Omega\Big)$$

where n is the number of grid cells in the study region.

All MCMC samplers used to fit model (2.1) were implemented from scratch in `R` version 3.4.3. For the simulations and analyses in this project computational runtime is not exorbitant, generally a matter of hours rather than days. However, future efforts could improve the speed of model fitting by implementing these samplers in `C`, `C++` or another more performative language.

Parameter Initialization

To facilitate convergence, Markov Chain Monte Carlo initial values for the parameters in model (2.1) were assigned in a special manner, taken as the output of simpler heuristic models fit to the distribution of observation sites and case-control counts. Our parameter initialization strategy entailed first fitting a spatial process model to the spatial pattern of observation sites, thereby obtaining crude estimates of w, θ and the T matrix, which were subsequently used to initialize these parameters in the real model fitting process. Case and control counts for each species were then regressed on the climatic covariates $z_\lambda(x)$ along with the crude estimate of the spatial random effects w, yielding estimates and initial values of β_{ms} and α_{ms}. The additional effort of fitting these heuristic models, rather than the simpler approach of randomly initializing parameters, was crucial to the success of model (2.1). As was the case for the single-species preferential sampling model presented in Project 1, convergence of model (2.1) to the correct parameter values of w and α_{ms} hinged upon this initialization strategy. We now provide specific details of the models fit to yield initial parameter values.

The following logistic regression model with a structural component logit-linear in spatially correlated random effects was fit to the observed pattern of sample sites for each species:

$$\kappa_{is}|\xi_{is} \sim \text{Bernoulli}(\xi_{is})$$
$$\text{logit}(\xi_{is}) = w_s(x_i)$$
$$w = (w_1, \ldots, w_s)^T, w \sim \text{MVGP}(0, H \otimes T)$$
$$H_{ij} = k(x_i, x_j; \theta)$$

Here, $\kappa_{is} \in \{0, 1\}$ denotes whether the ith grid cell of the study region contains a sampling event targeting species s, $w_s(x_i)$ is the value of a spatial random effect at the center point of the ith grid cell, x_i, for species s. Random effects for each species s are collected into vectors w_s, consisting of as many random effects as there are cells in the study region. The vector of all species-specific random effects, $w = (w_1, \ldots, w_s)^T$, follows a separable multivariate Gaussian process model. We note here that the model provided above is simply identical to the first 4 lines of model (2.1), stripped of those lines which describe case and control counts.

Having obtained crude estimates, and hence, initial values, for w, θ and T from the above model, case and control counts for each species at observed sites are regressed on the fixed effect covariates $z_\lambda(x)$ of model (2.1) along with the crude estimate of w, denoted $\hat{w}$, as follows:

$$Y_{ims}|w_s(x_i) \sim \text{Poisson}(\lambda_{ms}(x_i))$$
$$\log(\lambda_{ms}(x)) = z_\lambda(x)^T \beta_{ms} + \alpha_{ms} \times \hat{w}_s(x)$$

where Y_{ims} is the count in the ith observed grid cell for cases or controls, indexed by $m \in \{-, +\}$, for species s. That is, each species and disease status receives its own Poisson regression model, independent from those of other species and disease statuses. These regression models thus yield estimates and initial values for β_{ms} and α_{ms}, allowing model fitting to proceed.

Spatial Downscaling

We recall that model (2.1) entails discretizing the study region into equally sized, non-overlapping grid cells. Of course, for the practical purpose of producing disease risk maps is desirable for this resolution to be as fine as possible. However, the degree of resolution is ultimately limited by the computational burden of fitting the multivariate Gaussian process, w. For a study region consisting of n grid cells, and s total species, the covariance matrix of w is calculated as $\Sigma_w = H \otimes T$, where H is an $n \times n$ matrix containing the spatial decay in associations between grid cells, with the ijth element calculated as $H_{ij} = k(x_i, x_j; \theta)$ for correlation function $k(x_i, x_j; \theta)$, and T is the $s \times s$ cross-correlation matrix. Consequently, fitting model (2.1) requires inversion of the $n \times n$ spatial decay matrix H, not simply once, but at each draw of an MCMC sample. This computational burden quickly becomes prohibitive as n increases. For instance, if the study region considered was the state of California at a desired resolution of 16 km^2 then the covariance matrix would have 25,701 rows, or 660,541,401 elements, an intractable size.

Our strategy here is the same as that in Project 1, relying on the technique of spatial downscaling via thin plate spline interpolation. First, we fit model (2.1) at a lower resolution where convergence may be obtained after a reasonable period of time. This step yields coarse grain estimates of the spatial random effects $w = (w_1, \ldots, w_s)$, which consists of s different $n \times n$ vectors of random effects for each species. We

then separately interpolate each species-specific vector w_i $(i = 1, \ldots, s)$ to a high, fine grained resolution using thin plate splines, denoting the downscaled vectors as $\tilde{w}_s$, and our other low resolution parameter estimates as $\hat{\beta}_{ms}$ and $\hat{\alpha}_s$. To calculate the risk surface for species s, we make the assumption that the covariate relationships β_{ms} and α_s estimated at low resolution also hold at high resolution. Consequently, using the values of covariates $z_\lambda(x)$ known at high resolution, we can calculate log disease odds for point x as

$$z_\lambda(x)^T \hat{\beta}_{+,s} + \hat{\alpha}_{+,s} \times \tilde{w}_s(x) - z_\lambda(x)^T \hat{\beta}_{-,s} + \hat{\alpha}_{-,s} \times \tilde{w}_s(x)$$

which then easily yield the risk for that point in space.

2.3 Simulation 1: Comparative Performance

2.3.1 Introduction

A common theme of species distribution modeling is to share information across species in a joint analysis in order to obtain a better model for each species than what could be obtained from separate analyses. Key to this approach is the idea of exploiting correlations between species to correct for observational deficiencies. For instance, if there are particular areas of the study region where samples have been conducted for one species, but not another, then, provided there is some stochastic relationship between the two species, it is helpful to use information from the well observed species to influence predictions made about the other species in the areas for which it is sparsely observed. The proposed method of project 2 seeks to extend this approach to the context of preferential sampling in zoonotic disease surveillance.

The purpose of this simulation study is to assess the performance of the proposed multispecies model in comparison to the single species model of Project 1, in terms of reduced bias or reduced posterior variance in predicted risk.

In this study we consider the hypothetical scenario of a disease surveillance system targeting disease in two preferentially sampled species over the study region of California. We compare the proposed method (2.1) against two alternate approaches, one applying the single species model (1.1) of Project 1 to each species separately, and the other treating each species as indistinguishable, applying model (1.1) to the pooled data from both simulated species. We refer to the three approaches compared here as: 1) MVGP, for multivariate Gaussian process, referring to the proposed model (2.1) of the second project 2) Separate and 3) Pooled. We compare the performance of these approaches over 3 datasets simulated at different levels of inter-species correlation.

2.3.2 Data

The study region of California was discretized into a grid containing 458 equally sized cells, at resolution of 1,020 km^2. The first 2 principal components of the principal component decomposition of the 30 year normal PRISM climatic data were used as covariates $z_\lambda(x)$ of model (2.1), in addition to an intercept.

For this simulation study, 3 datasets were generated at differing values of the T matrix in model (2.1), intended to represent increasing strengths of inter-species correlation in terms of the latent processes w_1, w_2. These values are

$$T = \begin{bmatrix} 8 & 0 \\ 0 & 9 \end{bmatrix}$$

$$T = \begin{bmatrix} 8 & 3 \\ 3 & 9 \end{bmatrix}$$

$$T = \begin{bmatrix} 8 & 6 \\ 6 & 9 \end{bmatrix}$$

which we refer to as levels of none, medium and high inter-species correlation. Here, the off diagonal elements represent the marginal covariance in the random effect values between the species, as distance between measurements approaches zero. At each level of correlation, case and control counts for a total of $S = 2$ species were simulated from model (2.1), with values for the remaining parameters at each level provided in Table 2.1. For all levels of correlation, the range parameter of the exponential correlation function $k(x_i, x_j; \theta)$ was set to $\theta = 6$.

Correlation	Parameter	Value (Species 1)	Value (Species 2)
none	α_+	0.5	1
none	α_-	-0.5	-1
none	β_+	(-0.25, 0.75, -0.5)	(2, 0.8, -0.5)
none	β_-	(3.5, 0.5, 0.5)	(2.5, 0.5, 0.5)
medium	α_+	0.5	1
medium	α_-	-0.5	-1
medium	β_+	(1, 0.75, -0.5)	(2.25, 0.8, -0.5)
medium	β_-	(3.5, 0.5, 0.5)	(2.5, 0.5, 0.5)
high	α_+	0.5	1
high	α_-	-0.5	-1
high	β_+	(1.5, 0.75, -0.25)	(0.15, 0.5, -0.5)
high	β_-	(2.5, 0.5, 0.5)	(2.75, 0.5, 0.5)

Table 2.1: Parameters used to simulate data from the multispecies model at differing levels of inter-species correlation. The Correlation column refers to the magnitude of the off diagonal elements of the T matrix from the multivariate Gaussian process.

Case and control counts, along with prevalences, for each simulated species at each correlation level are summarized in Table (2.2).

Correlation	Species	Cases	Controls	Prevalence
none	1	1528	4927	0.237
none	2	2114	7989	0.209
medium	1	2319	8372	0.217
medium	2	2645	11862	0.182
high	1	1053	3729	0.220
high	2	1166	3315	0.260

Table 2.2: Case and control counts for each simulated species at each level of inter-species correlation.

Lastly, the distributions of observation sites for each level of correlation are summarized in figures 2.3, 2.4, and 2.5.

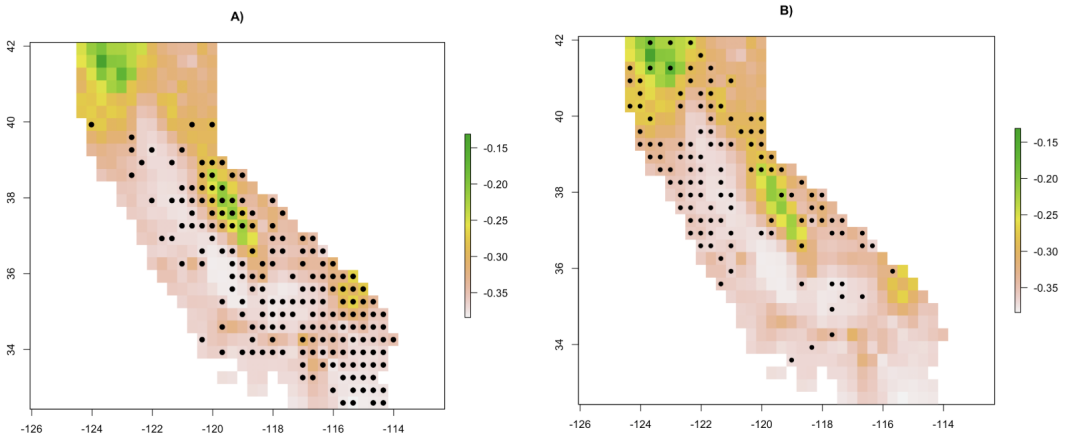

Figure 2.3: Distribution of observation sites under no inter-species correlation for A) species 1 and B) species 2. Black circles represent the location of an observation site.

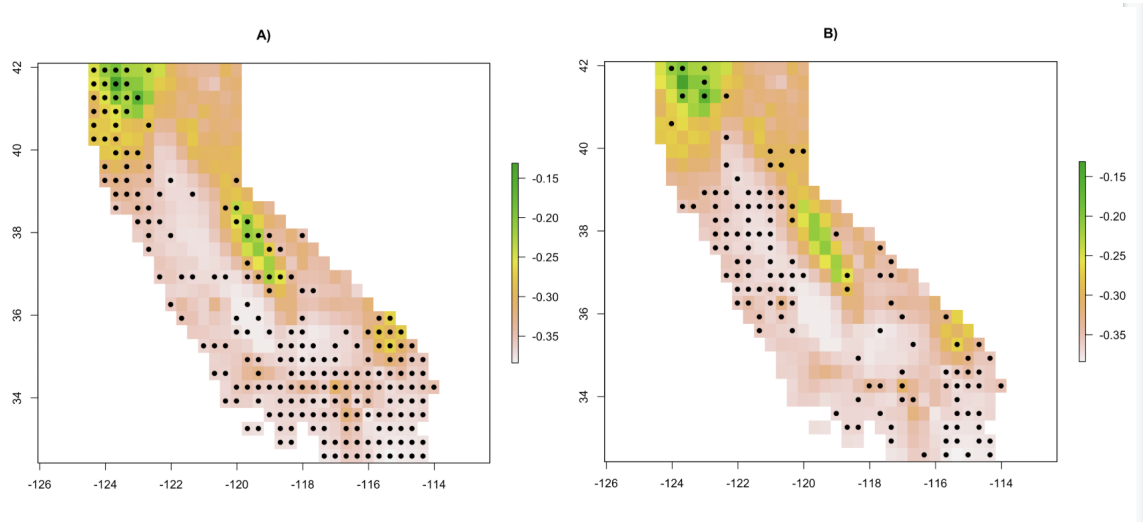

Figure 2.4: Distribution of observation sites under medium inter-species correlation for A) species 1 and B) species 2. Black circles represent the location of an observation site.

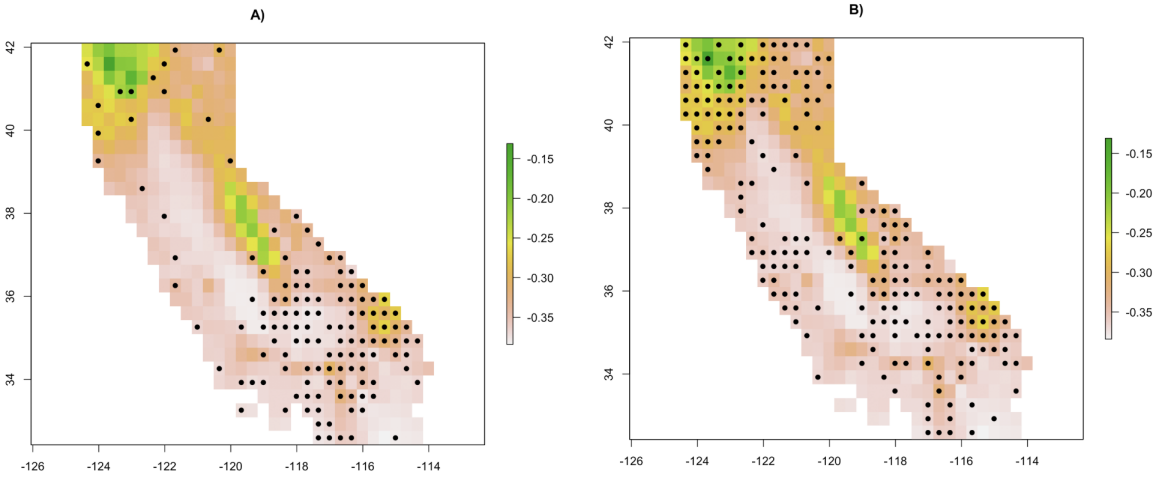

Figure 2.5: Distribution of observation sites under high inter-species correlation for A) species 1 and B) species 2. Black circles represent the location of an observation site.

2.3.3 Results

For each of the three simulated datasets a total of three different modeling approaches are compared, which we refer to as the MVGP (for multivariate Gaussian Process), separate and pooled approaches. The MVGP approach fits the multispecies model (2.1) to the surveillance data from both simulated species jointly, while the separate approach applies the proposed model from Project 1 to the data from each species separately. The pooled approach treats each species as identical, and applies the model from Project 1 to the combined data from both species. These three approaches are compared primarily through the metrics of root mean squared error (RMSE) in estimated disease log odds, and the magnitude of posterior variance in predicted disease risk. We first review the comparative RMSEs before delving into posterior variances on a dataset by dataset basis.

For all three levels of inter-species correlation, and for both simulated species, the pooled model showed the highest root mean squared error in predicted log disease odds (Table 2.3). However, the RMSEs of the pooled model decreased as the correlation level increased, dropping from (3.271, 4.000) for (species 1, species 2) under no correlation, to (2.081, 2.53) under medium correlation, to (1.366, 2.02) under high correlation. However, despite this drop, the MVGP model maintained a lower RMSE than the pooled model at all levels of correlation, The MVGP approach yielded slightly lower RMSEs than the separate model for both species under high correlation, and slightly lower RMSE under medium correlation for the second species.

Correlation	Species	Model	RMSE
none	1	MVGP	0.942
none	1	Separate	0.837
none	1	Pooled	3.271
none	2	MVGP	1.883
none	2	Separate	1.709
none	2	Pooled	4.00
medium	1	MVGP	0.758
medium	1	Separate	0.622
medium	1	Pooled	2.081
medium	2	MVGP	1.615
medium	2	Separate	1.744
medium	2	Pooled	2.53
high	1	MVGP	0.661
high	1	Separate	0.831
high	1	Pooled	1.366
high	2	MVGP	1.25
high	2	Separate	1.302
high	2	Pooled	2.02

Table 2.3: Root mean squared error in estimated log disease odds under different inter-species correlation levels.

We now conduct a deeper inspection of results for each level of correlation separately, considering in particular posterior variance in predicted disease risk and the distribution of errors in predicted disease log odds.

No Correlation

The first result set we further examine pertains to the case of no inter-species correlation, corresponding to the T matrix

$$T = \begin{bmatrix} 8 & 0 \\ 0 & 9 \end{bmatrix}$$

which gives the covariance between values of the random effects w in model (2.1) for

each species as distance reduces to zero. Posterior variance in predicted risk showed moderate differences across the three modeling approaches for both simulated species (Figure 2.6). When we examine the rasters of posterior variances from each approach, we first notice that the pooled model tends to reach a lower maximum variance, less than 0.1, for both species in comparison to the MVGP and Separate approaches, while the MVGP has overall lower variance than the Separate model. Secondly, for all three approaches, the first species shows a higher maximum variance, while at the same time, less overall posterior variance in regions outside those of the highest variances.

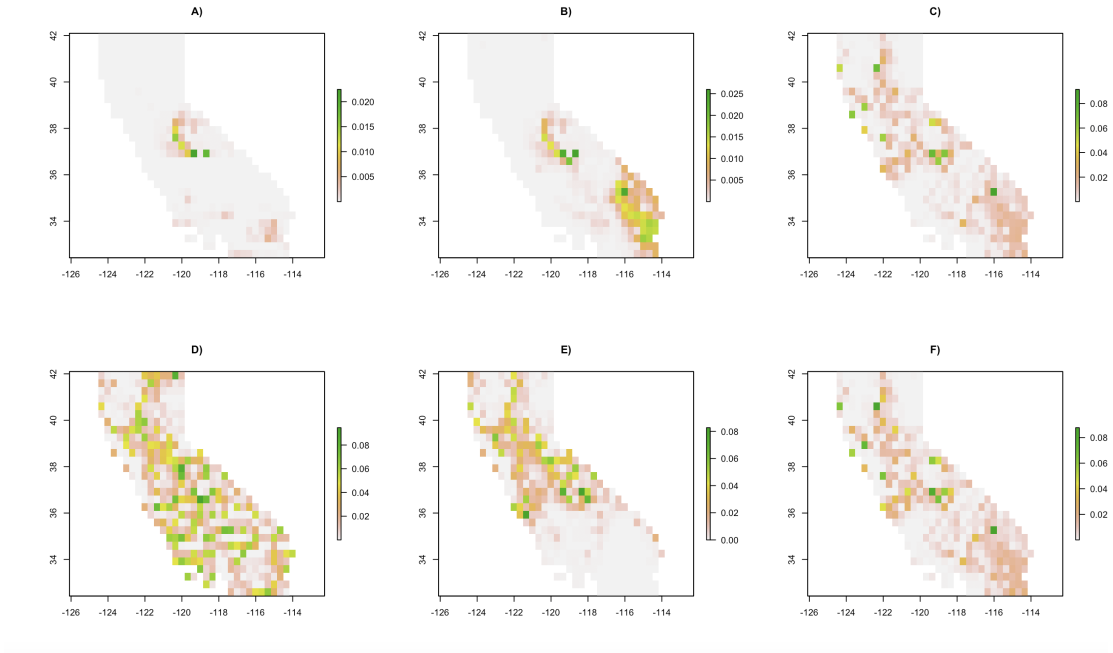

Figure 2.6: Posterior variance in risk, no inter-species correlation. A) species 1, mvgp model B) species 1, separate model C) species 1, pooled model D) species 2, mvgp model E) species 2, separate model F) species 2, pooled model

When we consider the distribution of error in predicted log disease odds, in the form of a scatterplot showing true log odds versus estimated log odds (Figure 2.7) for each cell in the study region, we see that the MVGP and Separate approaches show generally good agreement with the true log odds, having errors mostly evenly distributed above

and below the true log odds. For both approaches slight fanning out of error does occur as true log odds decrease, especially for the second simulated species, but the magnitude of this error is overall relatively contained. On the other hand, the pooled model shows a pronounced overestimation of log odds for several cells in the study region.

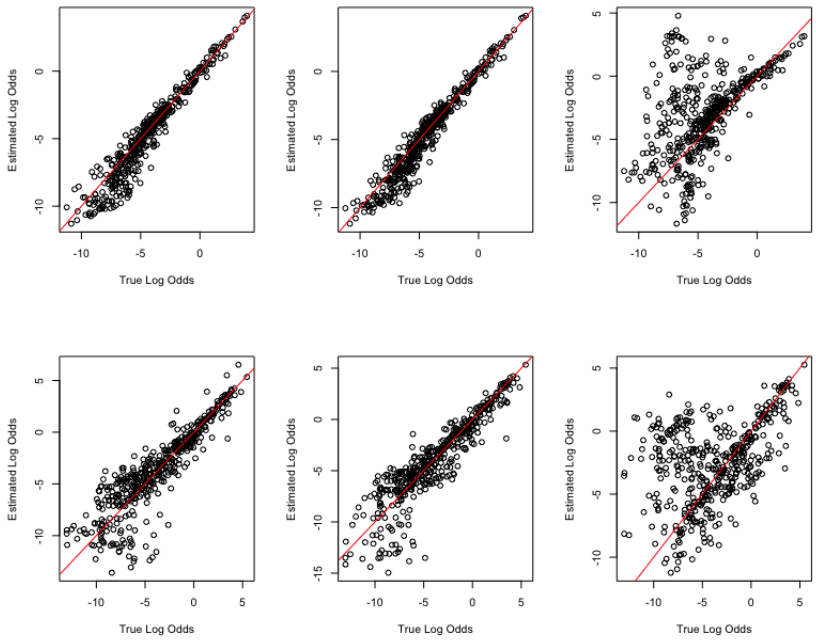

Figure 2.7: Log disease odds scatterplots with no correlation between the Gaussian processes of the simulated species. The first row corresponds to the estimated disease log odds of the first simulated species, and the second row to that of the second species. Within each row, the corresponding models for the scatterplots are, from left to right: multivariate Gaussian process model, the separate species model, and the pooled model.

Parameter estimates showed little differences as well between the MVGP and Separate models, both in terms of slopes and intercepts of β_+ and β_-, in addition to preferential sampling parameters α_+ and α_- (Table 2.4), (Table 2.5). Parameter biases and posterior variances remained low for these models. In contrast, the Pooled model showed higher bias, particularly for the first simulated species

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	1	-0.328	0.078	0.012
Separate	$\beta_{0,+}$	1	-0.316	0.066	0.020
Pooled	$\beta_{0,+}$	1	0.202	-0.452	0.011
MVGP	$\beta_{1,+}$	1	0.859	-0.109	0.003
Separate	$\beta_{1,+}$	1	0.850	-0.100	0.003
Pooled	$\beta_{1,+}$	1	0.749	0.001	0.003
MVGP	$\beta_{2,+}$	1	-0.527	0.027	0.002
Separate	$\beta_{2,+}$	1	-0.543	0.043	0.002
Pooled	$\beta_{2,+}$	1	-0.399	-0.101	0.003
MVGP	$\beta_{1,-}$	1	3.529	-0.029	0.004
Separate	$\beta_{1,-}$	1	3.501	-0.001	0.006
Pooled	$\beta_{1,-}$	1	3.858	-0.358	0.005
MVGP	$\beta_{2,-}$	1	0.383	0.117	0.002
Separate	$\beta_{2,-}$	1	0.392	0.108	0.002
Pooled	$\beta_{2,-}$	1	0.559	-0.059	0.002
MVGP	$\beta_{3,-}$	1	0.497	0.003	0.002
Separate	$\beta_{3,-}$	1	0.508	-0.008	0.002
Pooled	$\beta_{3,-}$	1	0.442	0.058	0.002
MVGP	α_{+}	1	0.523	-0.023	0.002
Separate	α_{+}	1	0.563	-0.063	0.003
Pooled	α_{+}	1	0.909	-0.409	0.014
MVGP	α_{-}	1	-0.528	0.028	0.001
Separate	α_{-}	1	-0.558	0.058	0.002
Pooled	α_{-}	1	-0.805	0.305	0.009

Table 2.4: Parameter estimates pertaining to species 1 with no correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	2	2.007	-0.007	0.024
Separate	$\beta_{0,+}$	2	1.930	0.07	0.016
Pooled	$\beta_{0,+}$	2	0.202	1.798	0.011
MVGP	$\beta_{1,+}$	2	0.781	0.019	0.005
Separate	$\beta_{1,+}$	2	0.782	0.018	0.007
Pooled	$\beta_{1,+}$	2	0.749	0.051	0.003
MVGP	$\beta_{2,+}$	2	-0.610	0.110	0.007
Separate	$\beta_{2,+}$	2	-0.436	-0.064	0.005
Pooled	$\beta_{2,+}$	2	-0.399	-0.101	0.003
MVGP	$\beta_{0,-}$	2	2.508	-0.008	0.024
Separate	$\beta_{0,-}$	2	2.601	-0.101	0.015
Pooled	$\beta_{0,-}$	2	3.858	-1.358	0.005
MVGP	$\beta_{1,-}$	2	0.535	-0.035	0.003
Separate	$\beta_{1,-}$	2	0.525	-0.025	0.006
Pooled	$\beta_{1,-}$	2	0.559	-0.059	0.002
MVGP	$\beta_{2,-}$	2	0.649	-0.149	0.006
Separate	$\beta_{2,-}$	2	0.468	0.032	0.005
Pooled	$\beta_{2,-}$	2	0.442	0.058	0.002
MVGP	α_{+}	2	1.053	-0.553	0.005
Separate	α_{+}	2	0.845	-0.345	0.005
Pooled	α_{+}	2	0.909	-0.409	0.014
MVGP	α_{-}	2	-1.090	0.590	0.005
Separate	α_{-}	2	-0.868	0.368	0.003
Pooled	α_{-}	2	-0.805	0.305	0.009

Table 2.5: Parameter estimates pertaining to species 2 with no correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Model (2.1) tended to slightly underestimate elements of the T matrix, in particular, the off-diagonal covariance terms (Table 2.6). Posterior variances for the elements of T were elevated as well, from 1.225 to 5.637. The spatial range θ was estimated with minimal bias (-0.935) and posterior variance (1.583).

Correlation	Parameter	Estimate	Bias	Posterior Variance
none	T (1,1)	7.82	-0.18	5.637
none	T (1,2)	-2.164	-2.164	1.225
none	T (2,1)	-2.164	-2.164	1.225
none	T (2,2)	6.522	-2.478	3.846
none	θ	5.065	-0.935	1.583

Table 2.6: Parameter estimates of T and θ from model (2.1) under no inter-species correlation.

Medium Correlation

Under medium inter-species correlation, the multispecies model tended to result in lower per-cell posterior variances in comparison to that obtained from separate applications of model (1.1). For the first simulated species, both multispecies and separate approaches resulted in a handful of cells in the mid-central portion of the map with elevated posterior variances, slightly above 0.020. However, for the multispecies model, outside of this region posterior variance tended to drop off to low levels, at or below 0.005, whereas the bottom portion of the map from the separate approach shows higher posterior variance, close to 0.010, consistently. However, the spatial distribution of posterior risk differs little across the different approaches for the second simulated species.

The distributions of error in predicted log disease odds show patterns similar to those observed in the case of no inter-species correlation. For both the multispecies and separate models, error tends to be distributed evenly when examining the scatterplot of true versus estimated per-cell log disease odds (Figure 2.9), with error moderately

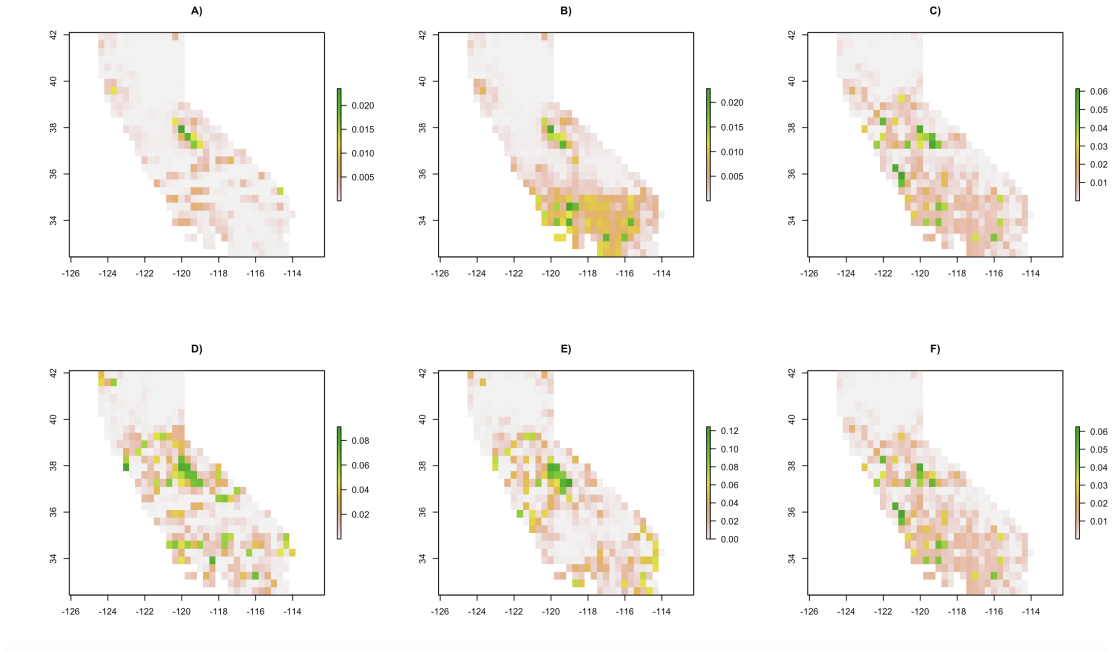

Figure 2.8: Posterior variance in estimated risk under medium inter-species correlation. A) species 1, mvgp model B) species 1, separate model C) species 1, pooled model D) species 2, mvgp model E) species 2, separate model F) species 2, pooled model

fanning out as true log odds decreases, especially for the second simulated species.

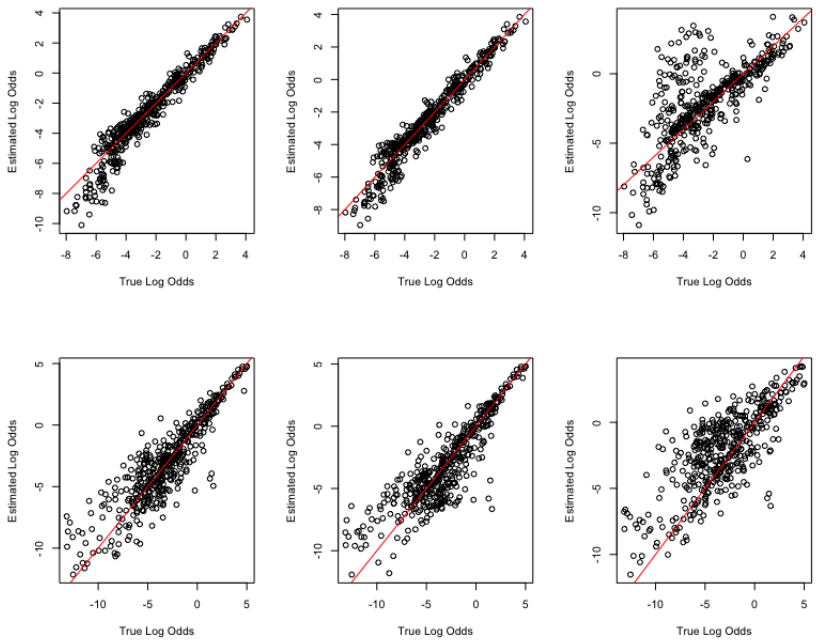

Figure 2.9: Log disease odds scatterplots with medium correlation between the Gaussian processes of the simulated species. The first row corresponds to the estimated disease log odds of the first simulated species, and the second row to that of the second species. Within each row, the corresponding models for the scatterplots are, from left to right: multivariate Gaussian process model, the separate species model, and the pooled model.

Parameter estimates remained in tandem for the MVGP and Separate approaches, in terms of bias and posterior variance. The pooled approach showed slightly greater bias for both simulated species, but to a lesser extent than witnessed under the case of no inter-species correlation in the previous section (Table 2.7), (Table 2.8).

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	1	0.982	0.018	0.004
Separate	$\beta_{0,+}$	1	0.992	0.008	0.007
Pooled	$\beta_{0,+}$	1	0.971	0.029	0.008
MVGP	$\beta_{1,+}$	1	0.804	-0.054	0.003
Separate	$\beta_{1,+}$	1	0.797	-0.047	0.003
Pooled	$\beta_{1,+}$	1	0.665	0.085	0.002
MVGP	$\beta_{2,+}$	1	-0.521	0.021	0.001
Separate	$\beta_{2,+}$	1	-0.515	0.015	0.001
Pooled	$\beta_{2,+}$	1	-0.517	0.017	0.003
MVGP	$\beta_{0,-}$	1	3.495	0.005	0.001
Separate	$\beta_{0,-}$	1	3.481	0.019	0.006
Pooled	$\beta_{0,-}$	1	3.900	-0.400	0.003
MVGP	$\beta_{1,-}$	1	0.468	0.032	0.003
Separate	$\beta_{1,-}$	1	0.477	0.023	0.003
Pooled	$\beta_{1,-}$	1	0.549	-0.049	0.001
MVGP	$\beta_{2,-}$	1	0.542	-0.042	0.002
Separate	$\beta_{2,-}$	1	0.538	-0.038	0.001
Pooled	$\beta_{2,-}$	1	0.540	-0.04	0.002
MVGP	α_{+}	1	0.510	-0.010	0.001
Separate	α_{+}	1	0.482	0.018	0.003
Pooled	α_{+}	1	0.690	-0.190	0.002
MVGP	α_{-}	1	-0.517	0.017	0.001
Separate	α_{-}	1	-0.486	-0.014	0.002
Pooled	α_{-}	1	-0.519	0.019	0.001

Table 2.7: Parameter estimates pertaining to species 1 with medium correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	2	2.306	-0.056	0.014
Separate	$\beta_{0,+}$	2	1.918	0.332	0.019
Pooled	$\beta_{0,+}$	2	0.971	1.279	0.008
MVGP	$\beta_{1,+}$	2	0.785	0.015	0.008
Separate	$\beta_{1,+}$	2	0.710	0.09	0.009
Pooled	$\beta_{1,+}$	2	0.665	0.135	0.002
MVGP	$\beta_{2,+}$	2	-0.475	-0.025	0.006
Separate	$\beta_{2,+}$	2	-0.552	0.052	0.009
Pooled	$\beta_{2,+}$	2	-0.517	0.017	0.003
MVGP	$\beta_{0,-}$	2	2.453	0.047	0.014
Separate	$\beta_{0,-}$	2	2.842	-0.342	0.020
Pooled	$\beta_{0,-}$	2	3.900	-1.400	0.003
MVGP	$\beta_{1,-}$	2	0.523	-0.023	0.007
Separate	$\beta_{1,-}$	2	0.601	-0.101	0.009
Pooled	$\beta_{1,-}$	2	0.549	-0.049	0.001
MVGP	$\beta_{2,-}$	2	0.469	0.031	0.006
Separate	$\beta_{2,-}$	2	0.553	-0.053	0.009
Pooled	$\beta_{2,-}$	2	0.540	-0.040	0.002
MVGP	α_{+}	2	1.154	-0.654	0.010
Separate	α_{+}	2	0.807	-0.307	0.002
Pooled	α_{+}	2	0.690	-0.19	0.002
MVGP	α_{-}	2	-1.156	0.656	0.008
Separate	α_{-}	2	-0.822	0.322	0.002
Pooled	α_{-}	2	-0.519	0.019	0.001

Table 2.8: Parameter estimates pertaining to species 2 with medium correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Bias in the spatial parameters of model (2.1), i.e. the T matrix and range parameter θ, was relatively contained but still noteworthy, at worst -4.105 for $T[2,2]$ (i.e. the second element on the diagonal of T) and at best -1.002 for $T[1,2]$ (Table 2.9). Posterior variances for the elements of T fell between 0.603 and 1.773.

Correlation	Parameter	Estimate	Bias	Posterior Variance
medium	T (1,1)	5.078	-2.922	1.773
medium	T (1,2)	1.998	-1.002	0.603
medium	T (2,1)	1.998	-1.002	0.603
medium	T (2,2)	4.895	-4.105	1.829
medium	θ	4.044	-1.956	1.466

Table 2.9: Parameter estimates of T and θ from model (2.1) under medium inter-species correlation.

High Correlation

Under high inter-species correlation the proposed method tended to result in lower posterior variance of predicted risk, compared to the separate applications of model (1.1), particularly for the first simulated species (Figure 2.10). For this species, under model (1.1) the map of posterior variances shows a narrow band of elevated variances towards the center of the map (Figure 2.10A), a band of much smaller area than the region of higher variance resulting from model (1.1) (Figure 2.10B), which covers the majority of the bottom portion of the map. The pooled model shows more extensive regions of higher variance than model (2.1) for the first species as well. For the second species, similar to the previous sections, all three modeling approaches tend to have higher posterior variances. However, the extent of the areas with higher posterior variance is still lesser for model (2.1) than for the other 2 approaches.

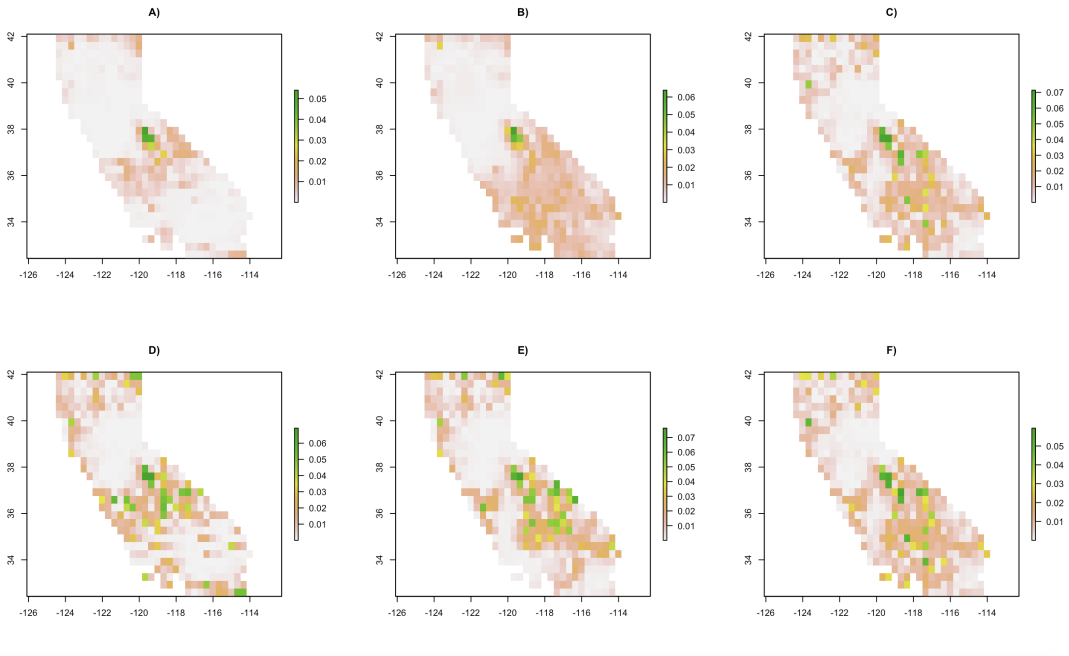

Figure 2.10: Posterior variance in risk, high inter-species correlation. A) species 1, mvgp model B) species 1, separate model C) species 1, pooled model D) species 2, mvgp model E) species 2, separate model F) species 2, pooled model

Scatterplots of true versus estimated log disease odds on for each grid cell in the study region (Figure 2.11) show good agreement for model (2.1) and the separate applications of model (1.1), with error for these approaches evenly distributed about the diagonal and only slightly fanning out as true log odds decreases. The pooled model showed much greater error, but less so than for the previous cases of no and medium inter-species correlation.

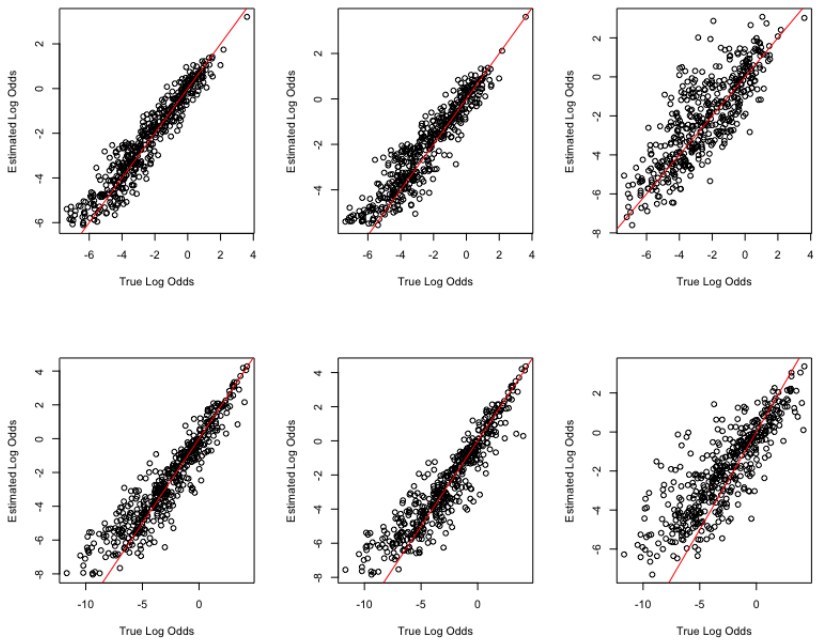

Figure 2.11: Log disease odds scatterplots with high correlation between the Gaussian processes of the simulated species. The first row corresponds to the estimated disease log odds of the first simulated species, and the second row to that of the second species. Within each row, the corresponding models for the scatterplots are, from left to right: multivariate Gaussian process model, the separate species model, and the pooled model.

Parameter estimates for both simulated species showed very contained posterior variances for all three approaches (Table 2.10), (Table 2.11). Similar estimates were obtained from the MVGP and separate applications of model (1.1), both of which had very little bias. The pooled model showed generally greater bias than the other two approaches, but less so than in the cases of no or medium inter-species correlation.

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	1	1.545	-0.045	0.010
Separate	$\beta_{0,+}$	1	1.627	-0.127	0.008
Pooled	$\beta_{0,+}$	1	0.737	0.763	0.007
MVGP	$\beta_{1,+}$	1	0.725	0.025	0.004
Separate	$\beta_{1,+}$	1	0.731	0.019	0.005
Pooled	$\beta_{1,+}$	1	0.537	0.213	0.005
MVGP	$\beta_{2,+}$	1	-0.266	0.016	0.004
Separate	$\beta_{2,+}$	1	-0.264	0.014	0.005
Pooled	$\beta_{2,+}$	1	-0.265	0.015	0.007
MVGP	$\beta_{0,-}$	1	2.536	-0.036	0.008
Separate	$\beta_{0,-}$	1	2.467	0.033	0.007
Pooled	$\beta_{0,-}$	1	3.476	-0.976	0.005
MVGP	$\beta_{1,-}$	1	0.508	-0.008	0.003
Separate	$\beta_{1,-}$	1	0.492	0.008	0.004
Pooled	$\beta_{1,-}$	1	0.448	0.052	0.006
MVGP	$\beta_{2,-}$	1	0.511	-0.011	0.003
Separate	$\beta_{2,-}$	1	0.516	-0.016	0.005
Pooled	$\beta_{2,-}$	1	0.578	-0.078	0.008
MVGP	α_{+}	1	0.447	0.053	0.002
Separate	α_{+}	1	0.401	0.099	0.002
Pooled	α_{+}	1	0.581	-0.081	0.004
MVGP	α_{-}	1	-0.478	-0.022	0.002
Separate	α_{-}	1	-0.441	-0.059	0.002
Pooled	α_{-}	1	-0.630	0.130	0.003

Table 2.10: Parameter estimates pertaining to species 1 with high correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Model	Parameter	Species	Estimate	Bias	Posterior Variance
MVGP	$\beta_{0,+}$	2	0.362	-0.212	0.012
Separate	$\beta_{0,+}$	2	0.316	-0.166	0.012
Pooled	$\beta_{0,+}$	2	0.737	-0.587	0.007
MVGP	$\beta_{1,+}$	2	0.554	-0.054	0.006
Separate	$\beta_{1,+}$	2	0.509	-0.009	0.014
Pooled	$\beta_{1,+}$	2	0.537	-0.037	0.005
MVGP	$\beta_{2,+}$	2	-0.469	-0.031	0.006
Separate	$\beta_{2,+}$	2	-0.472	-0.028	0.005
Pooled	$\beta_{2,+}$	2	-0.265	-0.235	0.007
MVGP	$\beta_{0,-}$	2	2.674	0.076	0.004
Separate	$\beta_{0,-}$	2	2.746	0.004	0.003
Pooled	$\beta_{0,-}$	2	3.476	-0.726	0.005
MVGP	$\beta_{1,-}$	2	0.419	0.081	0.006
Separate	$\beta_{1,-}$	2	0.467	0.033	0.017
Pooled	$\beta_{1,-}$	2	0.448	0.052	0.006
MVGP	$\beta_{2,-}$	2	0.519	-0.019	0.008
Separate	$\beta_{2,-}$	2	0.524	-0.024	0.005
Pooled	$\beta_{2,-}$	2	0.578	-0.078	0.008
MVGP	α_{+}	2	0.836	-0.336	0.008
Separate	α_{+}	2	0.758	-0.258	0.004
Pooled	α_{+}	2	0.581	-0.081	0.004
MVGP	α_{-}	2	-0.928	0.428	0.006
Separate	α_{-}	2	-0.857	0.357	0.003
Pooled	α_{-}	2	-0.630	0.130	0.003

Table 2.11: Parameter estimates pertaining to species 2 with high correlation between the Gaussian processes of the simulated species. Estimates are taken as posterior means.

Estimates of the spatial parameters from model (2.1), namely the spatial range θ and T matrix, showed modest degrees of bias, ranging from 2.458 to -1.552 for T and 1.927 for θ (Table 2.12), with moderate but not excessive posterior variances as well.

Correlation	Parameter	Estimate	Bias	Posterior Variance
high	T (1,1)	5.516	-2.484	4.462
high	T (1,2)	4.448	-1.552	2.664
high	T (2,1)	4.448	-1.552	2.664
high	T (2,2)	6.542	-2.458	4.479
high	θ	4.073	-1.927	1.776

Table 2.12: Parameter estimates of T and θ from model (2.1) under high inter-species correlation.

2.3.4 Discussion

This simulation study shed light on the performance of model (2.1), a hierarchical model containing a multivariate Gaussian process to capture correlation between species, relative to two alternative applications of model (1.1), referred to as the pooled and separate approaches. Model (2.1) and the separate application of model (1.1) greatly outperformed the pooled approach in terms of RMSE in predicted log disease odds, for all levels of inter-species correlation examined. The origin of this elevated error from the pooled approach is easy to pinpoint given that the pooled model treats each species as identical, whereas the true species specific climatic covariate values, $\beta_{+,s}$ and $\beta_{-,s}$, along with preferential sampling parameters $\alpha_{+,s}, \alpha_{-,s}$, and the random effects w_s differed by species. The pooled model thus amounted to making an incorrect modeling assumption regarding variability across species. However, had this assumption been correct and the relationships with climatic covariates and the sampling process been identical across species, we expect the pooled model to have performed better. This last point is evidenced by the fact that RMSE in log odds tended to decrease for the pooled model as inter-species correlation increased,

for the reason that as inter-species correlation increases, so too does the similarity between the simulated spatial random effects for each species, w_1 and w_2. For higher inter-species correlation levels, the proposed model (2.1) tended to have slightly lower RMSE than the separately modeled approach. However, the spatial coverage of the study region for both species is relatively ample (Figure 2.3), (Figure 2.4), (Figure 2.5), and consequently, it may be the case that the multispecies model would result in much lower error if at least 1 species were more sparsely observed.

In addition to this slight decrease in RMSE, the multispecies model also provided risk maps with slightly lower overall posterior variances. Generally speaking, compared to the separate applications of model (1.1), for medium and high levels of interspecies correlation, model (2.1) resulted in a lower maximum level of posterior variance in predicted risk, as well as a lesser spatial extent of area with higher posterior variance, especially for the first simulated species. Compared to the pooled approach the decrease in posterior variance of model (2.1) was still generally present, but not as pronounced as that for the separate analyses. However, the fact that the pooled model had such higher RMSE in predicted log odds than either of the two alternate approaches does not make the possibility of reduced posterior variance a compelling reason to use the pooled approach, given the model's problems with regard to RMSE.

One notable limitation of the proposed model (2.1) lies in the relationship between inter-species correlation and the spatial distribution of observation sites for each species. Recall that model (2.1) captures inter-species correlation specifically by way of correlation between random effect values for each species, denoted w_s, and mediated by the T matrix of the multivariate Gaussian process. Crucially, w_s values are shared between the components of the model describing case and control abundances, as well as those describing the distribution of sample sites. Consequently, as values of w_s become more correlated between different species, so too do the distributions

of observation sites for each species. This tendency may become problematic given that it would seem preferable to perform a joint, multispecies analysis when two or more species cover complementary domains of the study region. In such an instance, strengths in coverage of one species could be used to balance out the weaknesses in coverage of the others, and vice versa. However, correlation in w_s, and hence, correlation in the pattern of observation sites between species, may hinder model (2.1) from being applied to such a scenario. An alternate model that would constrain the locations of species less while still capturing inter-species correlation may be something akin to the following:

$$
\begin{aligned}
&Y_{ims}|w_s \sim \text{Poisson}(\lambda_{ims}) \\
&\log(\lambda_{ims}) = z(x_i)^T\beta_{ms} + \alpha_{ms} \times w_s(x_i) + \gamma_s(x_i) \\
&w_s \sim \mathcal{GP}(0, k(.,.;\theta_s, \phi_s)) \\
&\gamma = (\gamma_1, \ldots, \gamma_s)^T \sim \mathcal{MVGP}(0, \Sigma_\gamma) \\
&\kappa_{is}|p_{is} \sim \text{Bernoulli}(p_{is}) \\
&\text{logit}(p_{is}) = w_s(x_i)
\end{aligned}
$$

Here, instead of a multivariate Gaussian process as the shared latent process accounting for the effect of preferential sampling, the $w_s, s = 1, \ldots, S$ are now independent Gaussian processes, but the case and control rate functions λ_{ims} are also defined in terms of additional random effects as well, $\gamma_s(x_i)$, which originate from a multivariate Gaussian process. Here, the γ_s capture inter-species correlations but do not exert influence on the spatial distribution of sample sites, thus solving the initial problem. However there are a number of difficulties which complicate the implementation of this alternative model in practice. The added computational burden of multiple in-

dependent Gaussian processes w_s along with a multivariate Gaussian process γ_s is no small consideration. Moreover, the fact that each species is not necessarily observed at the same set of points prevents a separable covariance matrix from being specified for the γ_s terms, which complicates fitting of the multivariate Gaussian process. We leave pursuit of this model to future development.

2.4 Simulation 2: Model Robustness

2.4.1 Introduction

The proposed multispecies model of Project 2 hinges upon a key assumption, that of separability. We recall that model (2.1) attempts to correct for preferential sampling through the use of shared latent processes. In this framework, a multivariate spatial Gaussian process $w(s)$ is shared between the component of the model describing the distribution of sampling locations and that describing the abundance of cases and controls. In addition to adjusting for preferential sampling, the multivariate process also captures correlations in disease risk between species by assuming a separable covariance matrix, $H \otimes T$, where, for a discretized study region of n grid cells with S total species, H is and $n \times n$ matrix with elements $H_{ij} = k(x_i, x_j; \theta)$, describing the spatial decay in correlation between observations by way of parametric function $k(x_i, x_j; \theta)$, and T is an $S \times S$ matrix interpreted as the correlation matrix of the spatial random effects for all species at a particular point in space.

This assumption of separability, $H \otimes T$, constrains the spatial range of the Gaussian process for each species to be identical, which may be unreasonable in many real world disease surveillance applications. Thus, the purpose of this simulation study is to assess the performance of model (2.1) when provided with data which were not

generated according to a separable multivariate spatial process.

To that effect, when simulating data we replace the separable spatial process $w(x)$ defined above with a linear model of coregionalization. Linear models of coregionalization construct dependent, multivariate spatial processes as linear combinations of independent, univariate processes. For a full rank matrix A, and vector of independent, not necessarily identically distributed, spatial Gaussian processes $u(x)$, the technique of coregionalization constructs a multivariate spatial process $w(x)$ as

$$
\begin{aligned}
w(x) &= Au(x) \\
u(x) &= (u_1(x), u_2(x))^T \\
u_s(x) &\sim \mathcal{GP}(0, k_s(x, x'; \theta_i, \phi_i)) \text{ for } s \in \{1, 2\}
\end{aligned}
$$

where $k_s(x, x'; \theta_i, \phi_i)$ are known covariance functions of the Gaussian processes $u_s(x)$. Like the separable model, coregionalization models maintain the property of stationarity, specifying the correlation between observations solely as a function of their separation, rather than their absolute positions in space. However, key here is the ability of the coregionalization model to confer different spatial ranges to the component processes of $w(x)$, something not permitted by the separable model. We write the full, coregionalization based model from which we will simulate data as

$$
\begin{aligned}
&\kappa_{is}|\xi_s(x_i) \sim \text{Bernoulli}(\xi_s(x_i)) \\
&\text{logit}(\xi_s(x)) = w_s(x) \\
&w(x) = Au(x) \\
&u(x) = (u_1(x), u_2(x))^T \\
&u_s(x) \sim \mathcal{GP}(0, k_s(x, x'; \theta_i, \phi_i)) \text{ for } s \in \{1, 2\} \\
&Y_{ims}|w_s(x_i) \sim \text{Poisson}(\lambda_{ms}(x_i)) \\
&\log(\lambda_{ms}(x)) = z_\lambda(x)^T \beta_{ms} + \alpha_{ms} \times w_s(x)
\end{aligned} \tag{2.2}
$$

which is identical to model (2.1) except for the re-definition of $w(x)$ in terms of a coregionalization model. We will then fit model (2.1) to data simulated from model (2.2), equivalent to choosing a misspecified model. We shall evaluate the performance of model (2.1) in terms of the error in predicted log disease odds, along with a close inspection of the prediction error in the underlying spatial surface $w(x)$.

2.4.2 Data

The multispecies model (2.1) was fit to a total of 6 datasets, each of were simulated from the linear coregionalization model (2.2) and consisted of disease surveillance observations for 2 simulated species. These 6 simulated datasets were generated from a range of differing values for the matrix A, variances ϕ_i ($i \in \{1, 2\}$) and range parameters θ_i of model (2.2), with the intent being to evaluate performance of the proposed method under a wide variety of parameter specifications.

The first three simulated datasets specify the A matrix of the linear coregionalization

model as

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -0.5 \end{bmatrix}$$

and fix the marginal variance parameters of the Gaussian processes $u_1(x)$ and $u_2(x)$ as $\phi_1 = 1$, $\phi_2 = 1$. Spatial ranges θ_1 and θ_2 vary by dataset, take values of $(\theta_1, \theta_2) \in \{(2,4),(2,6),(5,15)\}$. The next three datasets fix the A matrix as

$$A = \begin{bmatrix} 0.25 & -1 \\ 0.25 & -0.5 \end{bmatrix}$$

and set greater marginal variances, $(\phi_1, \phi_2) = (3,3)$. Spatial range values increase following the same progression as the first three datasets, i.e. $(\theta_1, \theta_2) \in \{(2,4),(2,6),(5,15)\}$. For all simulated datasets the exponential covariance function was chosen for functions $k_s(x, x'; \phi, \theta)$ $(s \in \{1,2\})$ of the coregionalization model, which calculates the covariance between two points as

$$k(x, x'; \theta, \phi) = \phi \times \exp(-||x - x'||/\theta)$$

For each specification of coregionalization parameters, model (2.2) was simulated over the study region of California, which was discretized into 458 non-overlapping grid cells, at a resolution of 1090.3 km^2. The resulting number of observation sites for each species are summarized in Table (Table 2.13).

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	Observation Sites
(1, 1)	(2, 4)	1	175
(1, 1)	(2, 4)	2	192
(1, 1)	(2, 6)	1	137
(1, 1)	(2, 6)	2	197
(1, 1)	(5, 15)	1	96
(1, 1)	(5, 15)	2	125
(3, 3)	(2, 4)	1	91
(3, 3)	(2, 4)	2	142
(3, 3)	(2, 6)	1	87
(3, 3)	(2, 6)	2	126
(3, 3)	(5, 15)	1	52
(3, 3)	(5, 15)	2	110

Table 2.13: Numbers of simulated observation sites by species generated from the coregionalization model.

The first two principal components of the PRISM 30 year normal climatic dataset were used as covariates $z_\lambda(x)$ of the coregionalization model. For each dataset the remaining parameters of model (2.2), namely case and control climatic parameters β_{ms} associated with covariates $z_\lambda(x)$, and the preferential sampling parameters α_{ms}, were adjusted so as to yield disease prevalences between 0.13 and 0.23. These simulation parameters are summarized in Table (2.14), and the resulting disease prevalences in Table (2.15).

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	$\beta_{+,s}$	$\alpha_{+,s}$	$\beta_{-,s}$	$\alpha_{-,s}$
(1, 1)	(2, 4)	1	(1.25, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(1, 1)	(2, 6)	1	(2, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(1, 1)	(5, 15)	1	(2, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(1, 1)	(2, 4)	2	(1.75, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15
(1, 1)	(2, 6)	2	(1.75, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15
(1, 1)	(5, 15)	2	(1.75, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15
(3, 3)	(2, 4)	1	(2.2, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(3, 3)	(2, 6)	1	(2.45, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(3, 3)	(5, 15)	1	(2.65, 0.75, -0.5)	0.5	(3.5, 0.5, 0.5)	-0.25
(3, 3)	(2, 4)	2	(1.5, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15
(3, 3)	(2, 6)	2	(1.75, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15
(3, 3)	(5, 15)	2	(1.75, 0.8, -0.5)	0.45	(3.5, 0.5, 0.5)	0.15

Table 2.14: Additional simulation parameters used for the multispecies model robustness study.

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	Case Count	Control Count	Prevalence
(1, 1)	(2, 4)	1	1650	8536	0.16
(1, 1)	(2, 4)	2	2843	9721	0.23
(1, 1)	(2, 6)	1	1444	7803	0.16
(1, 1)	(2, 6)	2	2181	10028	0.18
(1, 1)	(5, 15)	1	980	5563	0.15
(1, 1)	(5, 15)	2	1410	5427	0.21
(3, 3)	(2, 4)	1	1386	8390	0.14
(3, 3)	(2, 4)	2	1493	6240	0.19
(3, 3)	(2, 6)	1	967	5685	0.15
(3, 3)	(2, 6)	2	1007	5939	0.14
(3, 3)	(5, 15)	1	553	3638	0.13
(3, 3)	(5, 15)	2	770	4542	0.14

Table 2.15: Summary of disease counts and prevalences by species simulated from the linear coregionalization model.

2.4.3 Results

The separable model (2.1) was fit to the 6 datasets simulated from model (2.2), a model incorporating linear coregionalization rather than a separable multivariate Gaussian process. In this section we evaluate model performance with respect to the error in predicted log disease odds and that in the estimated spatial random effects. For simplicity of presentation, we examine these results separately according to the A matrix and variance parameters (ϕ_1, ϕ_2) by which the data were generated. The first 3 datasets we review arise from specifying $(\phi_1, \phi_2) = (1, 1)$ and

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -0.5 \end{bmatrix}$$

while the next 3 datasets set $(\phi_1, \phi_2) = (3, 3)$ and

$$A = \begin{bmatrix} 0.25 & -1 \\ 0.25 & -0.5 \end{bmatrix}$$

For the first grouping, with $(\phi_1, \phi_2) = (1, 1)$, root mean squared error in log disease odds steadily increased as (θ_1, θ_2) increased, for both species (Table 2.16). Species 1 saw the RMSE take values of 1.18, 0.47, and 0.315 as (θ_1, θ_2) increased from $(2, 4)$ to $(2, 6)$ and then to $(5, 15)$. Overall species 2 had lower RMSE in log disease odds, which dropped from 0.282 to 0.15 to 0.122 as (θ_1, θ_2) followed that same progression of values.

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	RMSE
(1, 1)	(2, 4)	1	1.18
(1, 1)	(2, 6)	1	0.47
(1, 1)	(5, 15)	1	0.315
(1, 1)	(2, 4)	2	0.282
(1, 1)	(2, 6)	2	0.15
(1, 1)	(5, 15)	2	0.122

Table 2.16: Root mean squared error in predicted log disease odds by species when $(\phi_1, \phi_2) = (1, 1)$.

The distribution of error in predicted log disease odds (Figure 2.12) shows general directional agreement between true versus estimated values. Species 2 generally shows low levels of error at all values of (θ_1, θ_2), while species 1 shows greater error, especially as the value of the true log odds decreases. In particular, when $(\theta_1, \theta_2) = (2, 4)$, a notable segment of predicted values seems to underestimate the true log odds falling between -4 and -2.

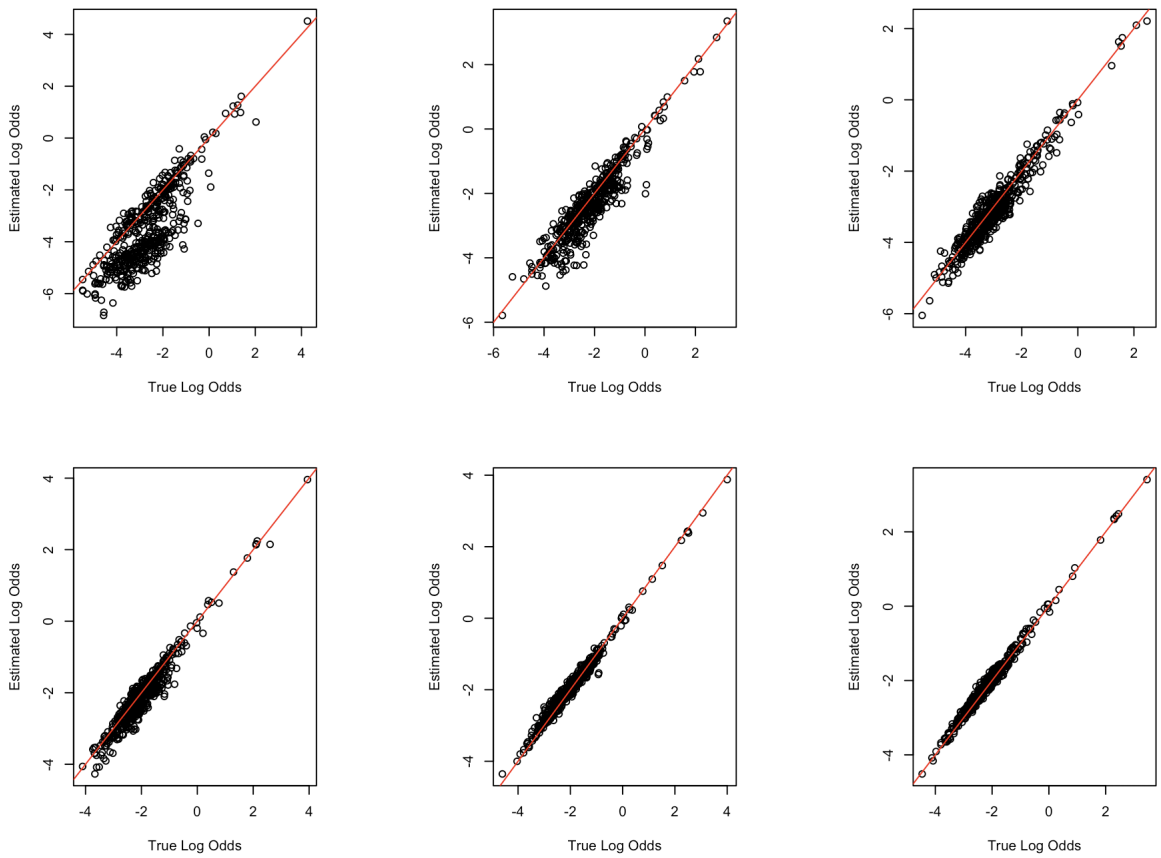

Figure 2.12: True versus estimated log disease odds when $(\phi_1, \phi_2) = (1, 1)$. The top row corresponds to the log odds of species 1, and the bottom to that of species 2. Values of (θ_1, θ_2) vary by column, being (2, 4), (2,6), (5,15) from left to right.

Root mean squared errors in estimated spatial random effects are generally higher for both species than RMSEs in log disease odds, but show a similar pattern of decrease as values of (θ_1, θ_2) increase (Table 2.19). RMSE for species 1 falls from 3.089 to 0.643 to 0.476 as (θ_1, θ_2) progresses from $(2, 4)$ to $(2, 6)$ to $(5, 15)$. Species 2 shows slightly greater RMSE than species 1 for corresponding levels of (θ_1, θ_2), i.e. 3.842, 0.704, and 0.548 for (θ_1, θ_2) at $(2, 4)$, $(2, 6)$ and $(5, 15)$.

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	RMSE
(1, 1)	(2, 4)	1	3.089
(1, 1)	(2, 6)	1	0.643
(1, 1)	(5, 15)	1	0.476
(1, 1)	(2, 4)	2	3.842
(1, 1)	(2, 6)	2	0.704
(1, 1)	(5, 15)	2	0.548

Table 2.17: Root mean squared error in estimated spatial random effects $w(x)$ by species when $(\phi_1, \phi_2) = (1, 1)$.

The distribution of error in predicted spatial random effects varies considerably by value of (θ_1, θ_2) (Figure 2.13). For $(\theta_1, \theta_2) = (2, 4)$, both species show a strange pattern in which greater values of w tend to be overestimated while lesser values tend to be underestimated. This pattern disappears for $(\theta_1, \theta_2) = (2, 6)$ and $(5, 15)$, for which estimated values tend to be more directionally correct. However, slight overestimation of greater values and underestimation of lesser values persists for both species.

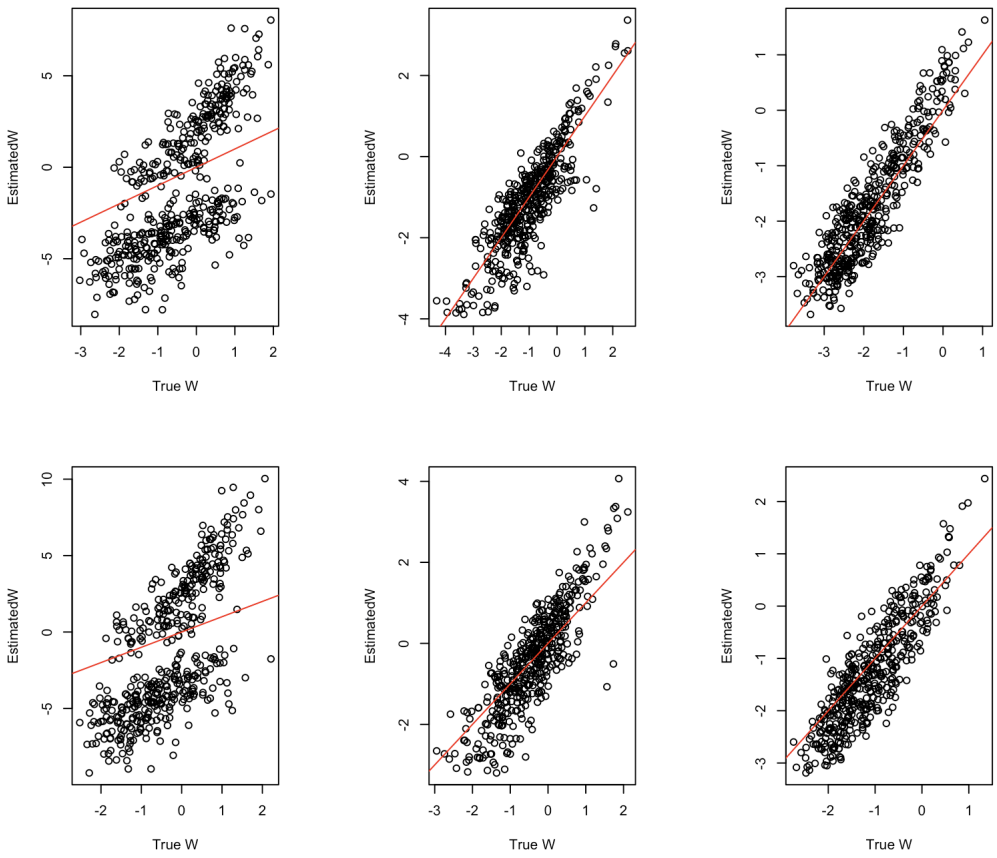

Figure 2.13: True versus estimated spatial random effects when $(\phi_1, \phi_2) = (1, 1)$. The top row corresponds to the log odds of species 1, and the bottom to that of species 2. Values of (θ_1, θ_2) vary by column, being (2, 4), (2,6), (5,15) from left to right.

We now turn our attention toward the final three datasets, which set $(\phi_1, \phi_2) = (3, 3)$ and

$$A = \begin{bmatrix} 0.25 & -1 \\ 0.25 & -0.5 \end{bmatrix}$$

Both species show slightly greater RMSE in predicted log disease odds than for the previous 3 datasets, where $(\phi_1, \phi_2) = (1, 1)$. For species 1, RMSE takes values of 1.188, 1.27, and 0.426 for (θ_1, θ_2) of $(2, 4)$, $(2, 6)$ and $(5, 15)$ (Table 2.18). RMSEs are

generally lower for species 2, taking values of 0.492, 0.265 and 0.114.

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	RMSE
(3, 3)	(2, 4)	1	1.188
(3, 3)	(2, 6)	1	1.27
(3, 3)	(5, 15)	1	0.426
(3, 3)	(2, 4)	2	0.492
(3, 3)	(2, 6)	2	0.265
(3, 3)	(5, 15)	2	0.114

Table 2.18: Root mean squared error in predicted log disease odds by species when $(\phi_1, \phi_2) = (3, 3)$.

Rather unusual patterns of error are apparent from an inspection of the scatterplots showing true versus estimated log disease odds (Figure 2.14). When $(\theta_1, \theta_2) = (2, 4)$ and $(2, 6)$, species 1 shows inflated error for true log odds less than 0, which falls both below and above the true log odds values, distributed in an unusual, asymmetric fashion. Species 2 shows a more consistent distribution of error, which tends to slightly overestimate the true log odds when $(\theta_1, \theta_2) = (2, 4)$ and $(2, 6)$. For both species the distribution of predicted values close in nicely to the true values for $(\theta_1, \theta_2) = (5, 15)$.

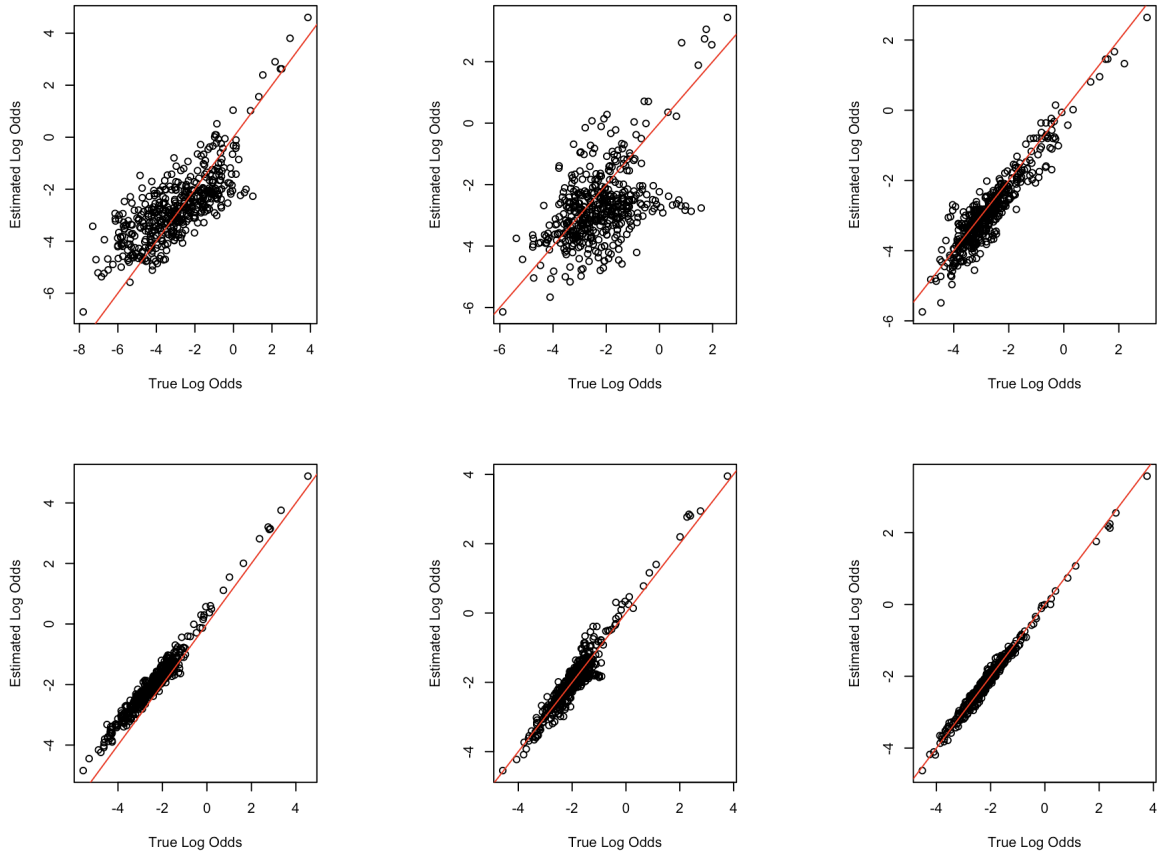

Figure 2.14: True versus estimated log disease odds when $(\phi_1, \phi_2) = (3, 3)$. The top row corresponds to the log odds of species 1, and the bottom to that of species 2. Values of (θ_1, θ_2) vary by column, being (2, 4), (2,6), (5,15) from left to right.

Root mean squared error in estimated spatial random effects decreased for both species as (θ_1, θ_2) increased (Table 2.19), dropping from 1.753 to 1.641 to 0.599 for species 1 and 1.243 to 1.142 to 0.383 for species 2.

(ϕ_1, ϕ_2)	(θ_1, θ_2)	Species	RMSE
(3, 3)	(2, 4)	1	1.753
(3, 3)	(2, 6)	1	1.641
(3, 3)	(5, 15)	1	0.599
(3, 3)	(2, 4)	2	1.243
(3, 3)	(2, 6)	2	1.142
(3, 3)	(5, 15)	2	0.383

Table 2.19: Root mean squared error in estimated spatial random effects $w(x)$ by species when $(\phi_1, \phi_2) = (3, 3)$.

The distributions of error in estimated spatial random effects w showed notably different patterns for these 3 datasets than for the previous 3 which set (ϕ_1, ϕ_2) to $(1, 1)$. When $(\theta_1, \theta_2) = (2, 4)$ and $(2, 6)$, both species showed widely spread errors for both positive and negative true values of w. However, when $(\theta_1, \theta_2) = (5, 15)$, the estimated values much more closely matched the true values (Figure 2.15).

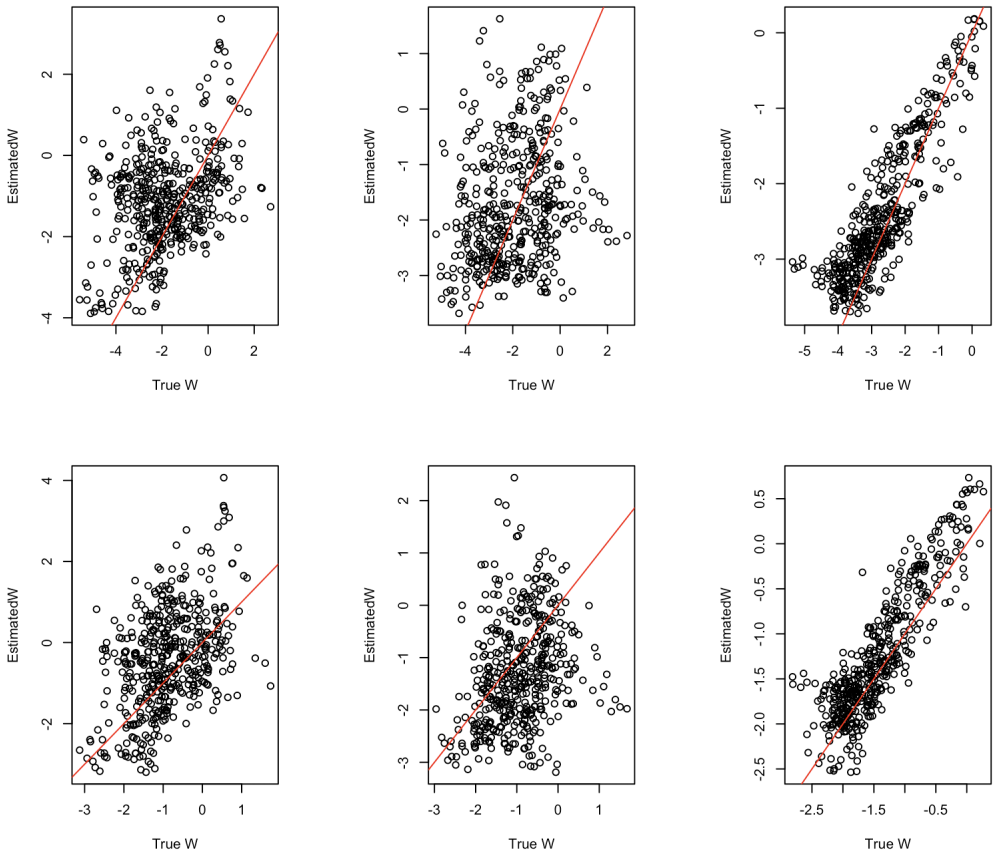

Figure 2.15: True versus estimated spatial random effects when $(\phi_1, \phi_2) = (3, 3)$. The top row corresponds to the log odds of species 1, and the bottom to that of species 2. Values of (θ_1, θ_2) vary by column, being (2, 4), (2,6), (5,15) from left to right.

2.4.4 Discussion

The purpose of this simulation study was to assess the performance of model (2.1) under circumstances where its assumption of separability fails to hold. To that effect, we simulated 6 datasets from an alternate model (2.2), relying on the technique of coregionalization to generate a multivariate spatial process, rather than the separable multivariate Gaussian process used in model (2.1). The results of this simulation study show that, when separability does not hold, the magnitude and distribution of error in the predicted log disease odds, as well as in the estimated spatial random

effects, is dependent upon the spatial ranges and marginal variances incorporated into the coregionalization model, being of high magnitude for some specifications of range and variance, and low, almost negligible, magnitude for others.

The first major point of discussion we raise is the fact that non-separability can result in poor accuracy in predicted log disease odds, and in the estimated spatial random effects, depending on the spatial processes which were used to construct the coregionalization model. We recall that coregionalization models a multivariate spatial process whose response is of dimension 2 as

$$
\begin{aligned}
w(x) &= Au(x) \\
u(x) &= (u_1(x), u_2(x))^T \\
u_s(x) &\sim \mathcal{GP}(0, k_s(x, x'; \theta_i, \phi_i)) \text{ for } s \in \{1, 2\}
\end{aligned}
$$

where A is a full rank matrix and $u_s(x)$ are Gaussian processes. The choice of parameters θ_i, ϕ_i in the covariance functions of these processes seemed to have the greatest impact on the magnitude of error in predicted log odds and estimated values of w. Of the different values of (θ_1, θ_2) considered, $(\theta_1, \theta_2) = (2, 4)$ resulted in the highest root mean squared error, both of log odds and w, followed by $(\theta_1, \theta_2) = (2, 6)$ and then $(\theta_1, \theta_2) = (5, 15)$. That is, error tended to decrease as the values of the θ_i increased, so much so that for $(\theta_1, \theta_2) = (5, 15)$ the error was of negligible magnitude. To understand this decrease, we return to the covariance function of the coregionalization model

$$k(x, x'; \theta, \phi) = \phi \times \exp(-||x - x'||/\theta)$$

We see that in this covariance function, as θ increases the covariance between two points x, x' decreases, and therefore, in general, incorrect estimation of the spatial structure becomes less consequential because the strengths of spatial associations are overall of lesser magnitude.

Another observation worthy of discussion is the fact that for most datasets, error in the log disease odds of species 2, the second simulated species, was moderately lower than that of species 1. To understand this distinction we re-examine Tables (2.15) and (2.13), and note that the total number of collected specimen (cases and controls) for species 2 was always greater than that of species 1, which, taken with the fact that species 2 also had more simulated locations than species 1, allowed for greater information in estimating the random effects and covariates for this species.

In conclusion, the simulations conducted in this study give a reasonable summary of the performance of model (2.1) when separability does not hold, showing that the error in predicted log disease odds and estimated spatial random effects can inflate for low values of the spatial range parameters. However, as the scales of the component processes decrease, the magnitude of this error is by no means catastrophic or massive, and the distribution of errors show that the predicted log disease odds are at least generally directionally correct (Figure 2.12, Figure 2.14). Thus, while not entirely robust to non-separability, the proposed model can still be moderately or even strongly performative, depending on the spatial properties of the underlying process.

One prominent limitation of this study is the narrow way in which violations of the assumption of separability were manifested. To break the assumption of separability spatial random effects were simulated from a coregionalization model, in which the spatial ranges of the component Gaussian processes differed. However, the input processes still were both of mean zero, while it is likely that specifying different,

nonzero means for these processes may have resulted in greater errors. Moreover, the resulting spatial process of the coregionalization model still maintained the property of stationarity, in which correlations between observations at different locations were a function of the separation of those locations, rather than their absolute locations in space. In real world disease surveillance applications monitoring multiple species, circumstances may arise wherein two species are highly correlated in one particular area of the study region, but not in others. Stationary models would fail to capture this distinction. To overcome this pitfall future efforts may be devoted to extending model (2.1) to encompass nonstationary spatial processes.

2.5 Analysis

2.5.1 Introduction

This analysis considers a disease surveillance application encompassing multiple species, with the intent to apply the joint model proposed in the methods section of Project 2. As in Project 1, the disease surveillance system of interest is operated by the California Department of Public Health (CDPH), targeting plague (infection by *Yersinia pestis*) among its animal hosts across the state of California. However, in this analysis we consider not only data pertaining to plague in Sciurids, the rodent family of squirrels, but also to coyotes, both of which are hosts for plague, and are monitored by preferential sampling mechanisms. The objective of analysis is thus to estimate plague risk maps for Sciurids and coyotes through a joint modeling framework which adjusts for preferential sampling. These last two points, joint modeling and preferential sampling, are key areas of focus in this application. Regarding the first, we wish to ascertain the strengths of correlations estimated between the two species groups,

which are good indicators of how much information can be "shared" between groups, and thus a reflection of any gain to be acquired by using a joint model as opposed to a single-species model. We emphasize that while we will ultimately produce separate risk maps for plague in coyotes and Sciurids, the map for each species grouping will in principle be able to borrow information from data from the other species. For the second point, we wish to characterize the degree or estimated effect of preferential sampling on the spatial predictions of risk.

The first species grouping monitored by the CDPH plague surveillance system is the Sciurid family, encompassing 21 different species of rodents in the observed data (i.e. Antelope Ground Squirrel, Antelope Ground Squirrel, Belding's Ground Squirrel, California Ground Squirrel, Chipmunk, Least Chipmunk, Long-eared Chipmunk, Lodgepole Chipmunk, Merriam's Chipmunk, Panamint Chipmunk, Shadow Chipmunk, Siskiyou Chipmunk, Sonoma Chipmunk, Uinta Chipmunk, Yellow-pine Chipmunk, Golden-mantled Ground Squirrel, Ground Squirrel, Yellow-bellied Marmot, Pine Squirrel, and Squirrel). The surveillance system collects data by conducting a series of sampling events at locations throughout California, in which Sciurids are trapped and subsequently tested for *Yersinia Pestis*. The data contain samples collected between 1983 and 2015. This analysis aggregates data for all Sciurid species, and for all years observed. The surveillance system predominantly collects data in a strategy of preferential sampling, assigning sampling locations to high risk or high impact areas. Here, risk is assessed to be high in what are viewed as plague endemic regions, as determined by by historic cases of plague in humans or recovered Sciurid specimens, and high impact areas are regions with high potential for human Sciurid interactions, such as in national parks or areas which are climatically suitable for plague and have high human usage, such as the Lake Tahoe area in the northeastern portion of the state.

The sampling mechanism for coyotes (*Canis latrans*) follows a different chain of events than that of the Sciurids but one which nevertheless can be viewed as a form of preferential sampling. In the first step of the sampling process, coyotes are recovered either in the form of roadkill or in response to livestock harassment reports. In either case, coyote carcasses are collected and submitted to the CDPH along with locational identifiers describing the point of recovery. These identifiers vary in precision from high quality, such as latitude and longitude coordinates obtained from GPS, to verbal directions, such as the estimated mileage from a nearby road or other identifier. We note here the potential for misclassification error in the latter form of description, due to the fact that some verbal directions may be incorrect, imprecise or notably different than the true point of recovery. Upon reception of a carcass, CDPH conducts F1 antigen blood tests for plague only if the carcass was recovered in what is deemed to be a plague endemic region. This final step, the conduction of tests conditional on origin from a plague endemic area, is a form of preferential sampling, as tests for plague are thus conducted only in at-risk areas. Consequently, the adjustment for preferential sampling undertaken in the proposed method is warranted when analyzing the coyote data as well, in addition to the preferentially sampled Sciurids.

To characterize the impact of the joint modeling structure of (2.1), we compare the Sciurid and coyote risk maps obtained from this model with those produced by the model introduced in Project 1.

By applying the Project 1 model to the Sciurid and coyote data separately, and comparing the resulting risk maps with those estimated by model (2.1), we primarily seek to ascertain whether the joint model produces quantitatively different estimates of risk compared against this baseline method, as well as whether the joint model results in risk estimates with lower posterior variance, as a result of "information sharing" between different species groups.

2.5.2 Data

For both species groupings, Sciurids and coyotes, the data consist of latitude and longitude coordinates pertaining to recovered specimens, along with species identifications, timestamps marking the date of observation and the results, either positive or negative, of F1 antigens test for plague. As was discussed above, for many of the recovered coyotes the raw data do not provide the exact geo-coordinates at which the animal was found, but rather offer verbal directions describing the site of recovery. In such cases the directions were manually converted to a set of latitude and longitude values, and a measure of confidence ranging from low to high was also recorded to indicate faith in these estimated coordinates. The extensive amount of effort involved in this manual conversion was undertaken by Dr. Ian Buller (2019), to whom we are deeply indebted. For the purpose of this analysis all recovered coyote observations whose geocoded confidence was not considered poor were used.

The plague surveillance data for Sciurids contains observations conducted between 1983 and 2015, while that for coyotes covers 1984 through 2015. Data from all available years were used for each species grouping, despite the 1 year difference in starting times, which we argue is negligible in comparison to the overall timespan of observation. The Sciurid data from 1983 were inspected and found not to contain any extreme outliers in terms of recorded disease prevalence or number of sampling sites, compared to all subsequent years. From 1983 to 2015 a total of 20,366 Sciurid specimen were recovered, with an overall disease prevalence of 0.064 (Table 2.20). For all available years 7,250 coyotes were recovered, 704 of which (8.9%) tested positive for plague.

Species	Cases	Controls	Prevalence
Sciurids	1401	20366	0.064
Coyotes	704	7250	0.089

Table 2.20: Summary of plague case and control counts for Sciurids and coyotes.

Sciurids were recovered from a total of 1,751 different observation sites throughout the state of California. The geo-coordinates of coyotes, either provided in the raw data or manually estimated from the recorded directions, cover a total of 4,318 different points in space or observation sites. The distributions of observation sites of both species groupings are plotted in Figure (2.16). We note here that the combined distribution of both species groupings offers a decisively greater coverage of the study region than what could be gained from any single grouping alone. Coyote observations provide better coverage of the coastal region and north eastern corner of the map, in addition to a small pocket along the southern border of the state near the Imperial Valley region, while Sciurid observations cover the eastern portion of the Sierra Nevada mountain range in much greater numbers. Leveraging this complimentary spatial coverage of the study region motivates the use of a joint modeling approach, in the hope that borrowing information from each species can yield better spatial predictions.

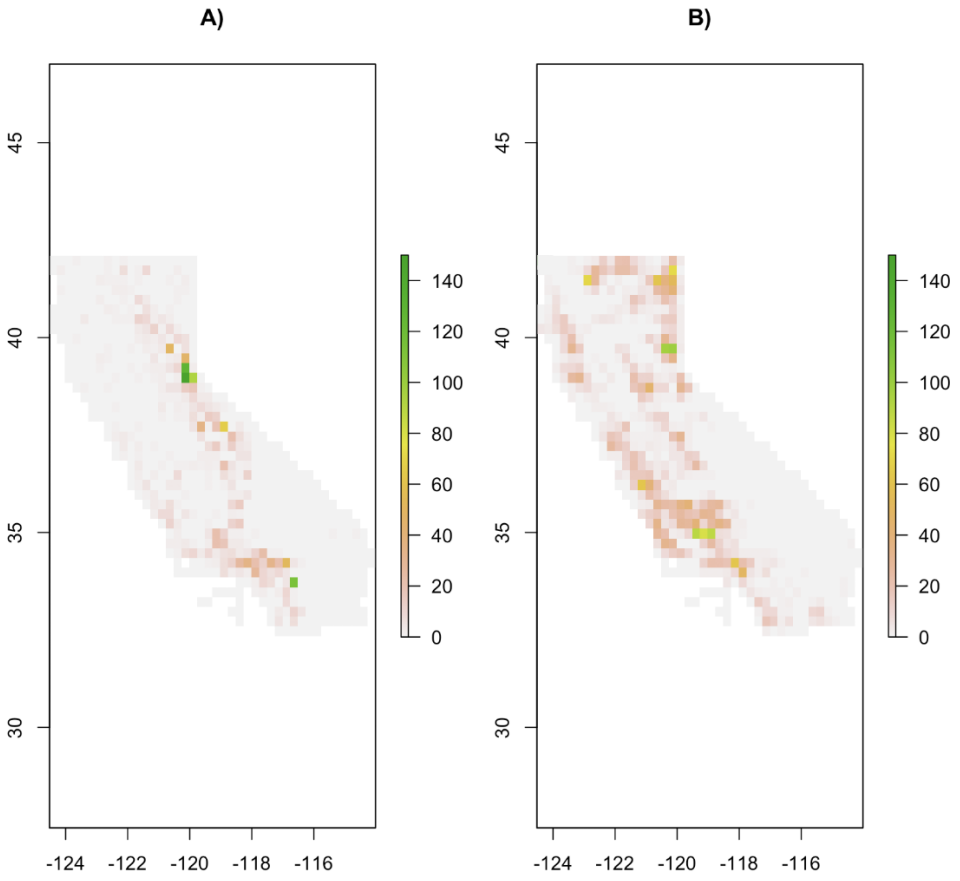

Figure 2.16: Distributions of distinct sampling locations for A) rodents and B) coyotes between 1982 and 2015.

Lastly, for the covariates $z_\lambda(x)$ in model (2.1) the first two principal components of the PRISM 30-year average climatic normal measurements were used (Figure 2.17). Specifically, principal components were calculated from 7 original climatic measurements, namely mean, minimum, and maximum temperature, mean dew point temperature, precipitation, and minimum and maximum vapor pressure deficits.

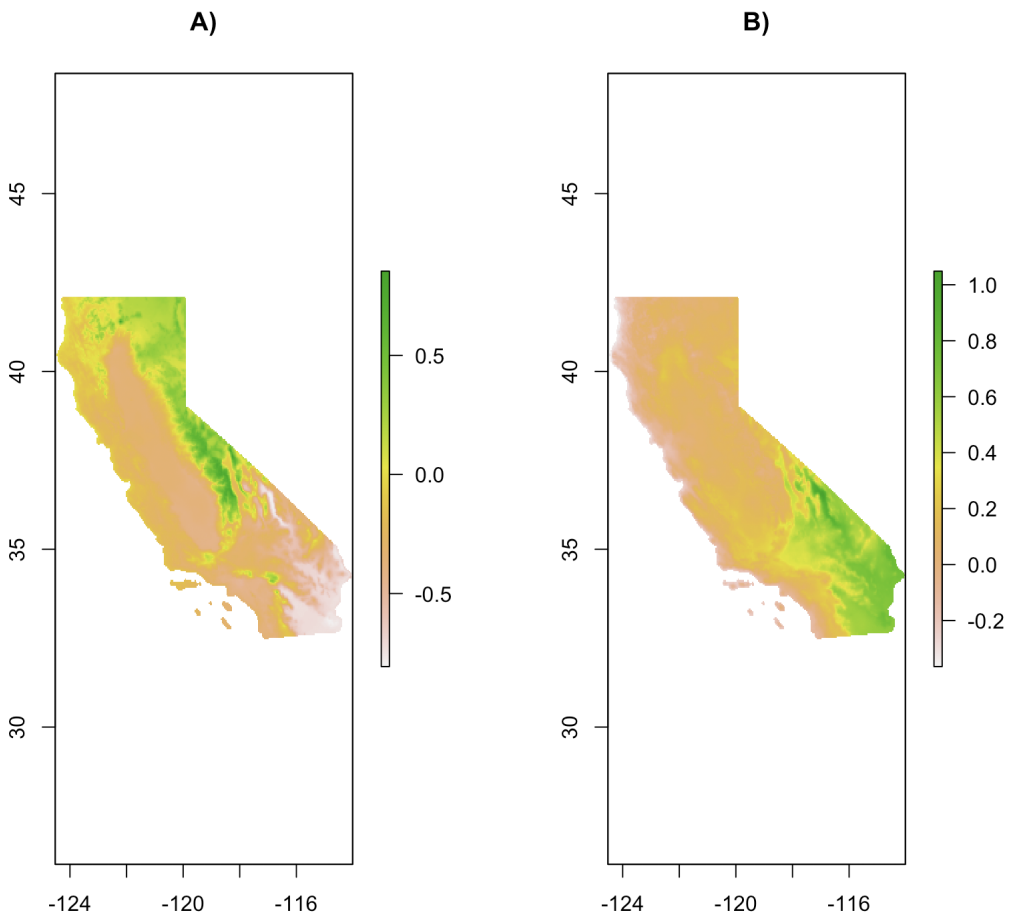

Figure 2.17: A) First and B) second principal components of the PRISM 30 year climatic normals used as covariates for the joint coyote-Sciurid analysis, at 16 km^2 resolution.

2.5.3 Results

Model (2.1) was fit jointly to the Sciurid and coyote disease surveillance data at a resolution of 614 km^2 via MCMC. Model convergence was ascertained through inspection of parameter traceplots, with a total of 5,000 MCMC samples drawn and a burnin of 1000. The estimated multivariate spatial process w was downscaled via thin plate spline interpolation to a resolution of 16 km^2, from which high resolution risk maps for plague in Sciurids and coyotes were obtained.

The risk map for plague in coyotes (Figure 2.18) shows a lengthy band of elevated risk

stretching along the Sierra Nevada mountain range, continuing up to the northeastern corner of the state, reaching a maximum risk value of 0.56. Two additional neighborhoods of elevated risk arise in southern California. In contrast, the San Joaquin Valley, to the west of the Sierra Nevada mountains, shows a near zero level of risk for plague in coyotes, while the coastline possesses low, but nonzero, risk.

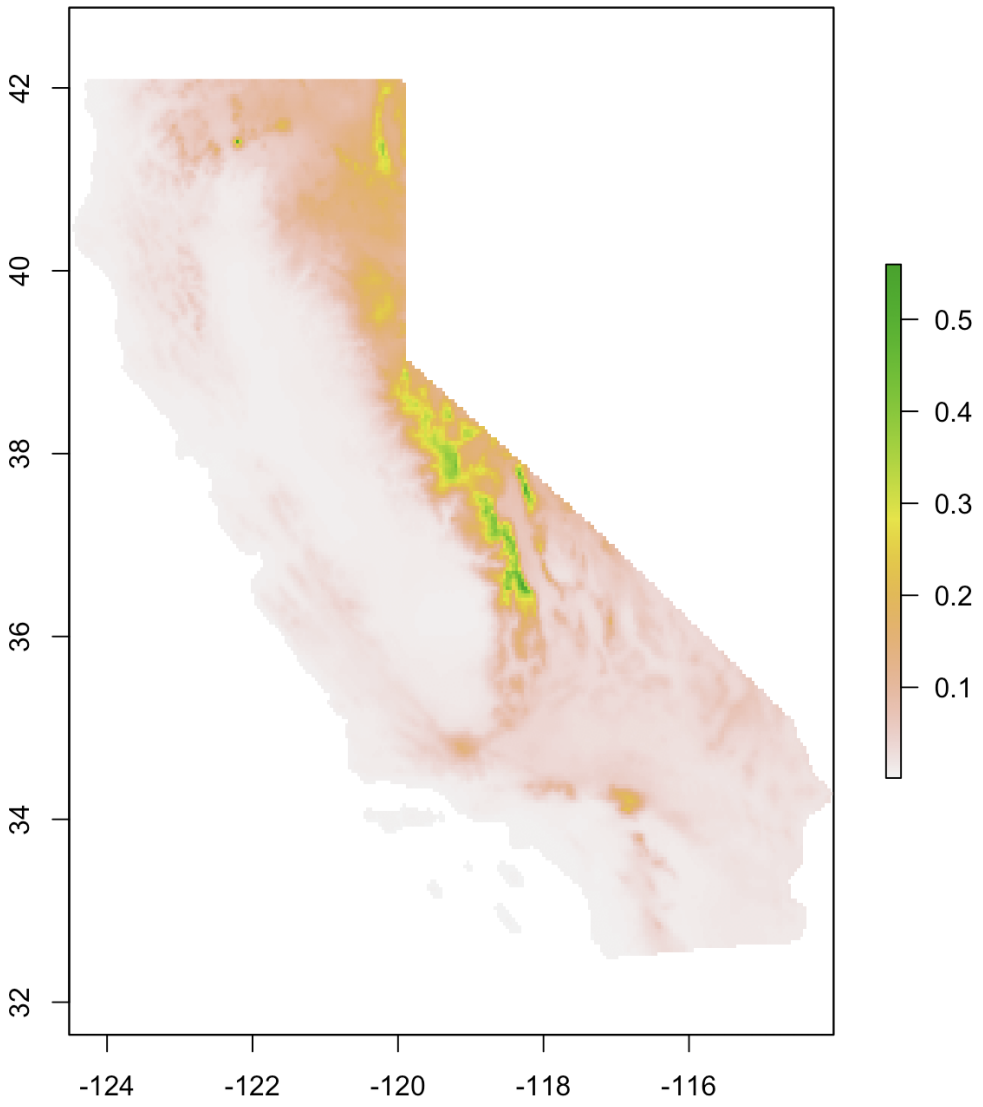

Figure 2.18: Coyote risk map at 16 km^2 resolution obtained from the multivariate Gaussian process model.

For a rough visual inspection of the influence of observed cases and controls on the resulting risk estimates, we juxtapose the estimated coyote risk map with rasters depicting the abundances of observed disease positive and negative coyotes (Figure 2.19). We see that in the areas of elevated risk in the northeastern corner of the state coincide with dense clusters of cases. Interestingly, the high risk areas towards the central and southern portion of the Sierra Nevada mountain range show few observed cases or controls. The 2 main pockets of cases in southern California are matched by a more slight elevation in risk, between 0.10 and 0.20.

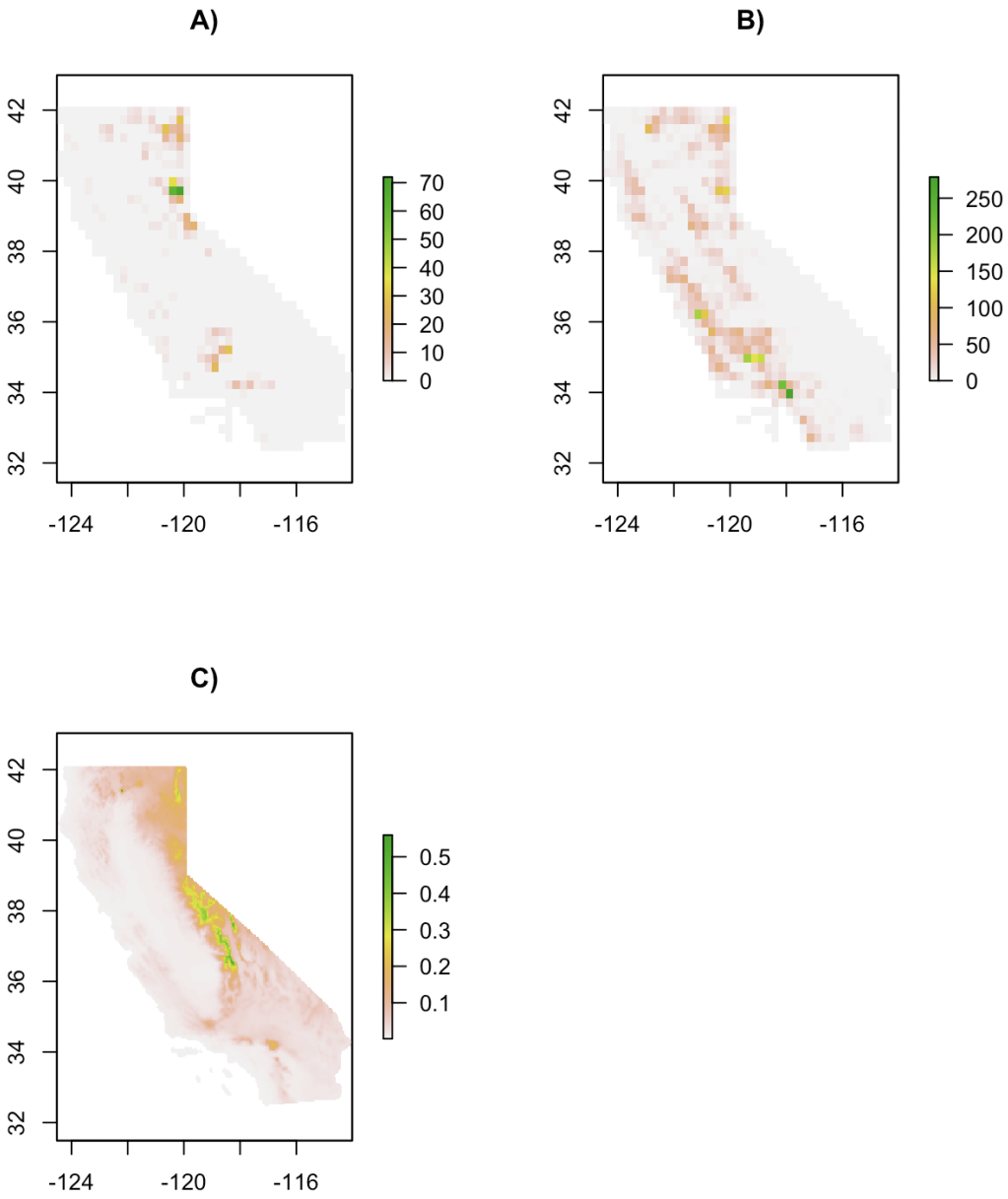

Figure 2.19: Distributions of A) case counts, B) control counts, and C) estimated risk for plague in coyotes between 1983 and 2015.

To help explain the areas of elevated risk in the coyote plague map, especially those in the places with few observed cases (principally the Sierra Nevada mountain range) we separate the contributions to predicted risk put forward by covariates and random effects. We recall from model (2.1) that the log disease odds for a particular point in space x are given by

$$z_\lambda(x)^T\beta_{+,s} + \alpha_{+,s} \times w_s(x) - z_\lambda(x)^T\beta_{-,s} + \alpha_{-,s} \times w_s(x)$$

We can separate this expression into firstly a covariate contribution, $z_\lambda(x)^T\beta_{+,s} - z_\lambda(x)^T\beta_{-,s}$, due to the PRISM climatic principal components used here as fixed effects $z_\lambda(x)^T$, and secondly the expression $\alpha_{+,s} \times w_s(x) - \alpha_{-,s} \times w_s(x)$ provided by the spatially structured random effects $w_s(x)$. Side by side inspection of these different contributions helps explain the areas of high risk as either due to covariates, or random effects, or a mixture of both (Figure 2.20). From this inspection, we see that the northeastern portion of the map demonstrating elevated risk has both high covariate and random effect contributions, while risk in the Sierra Nevada mountains is almost entirely driven by covariate contributions.

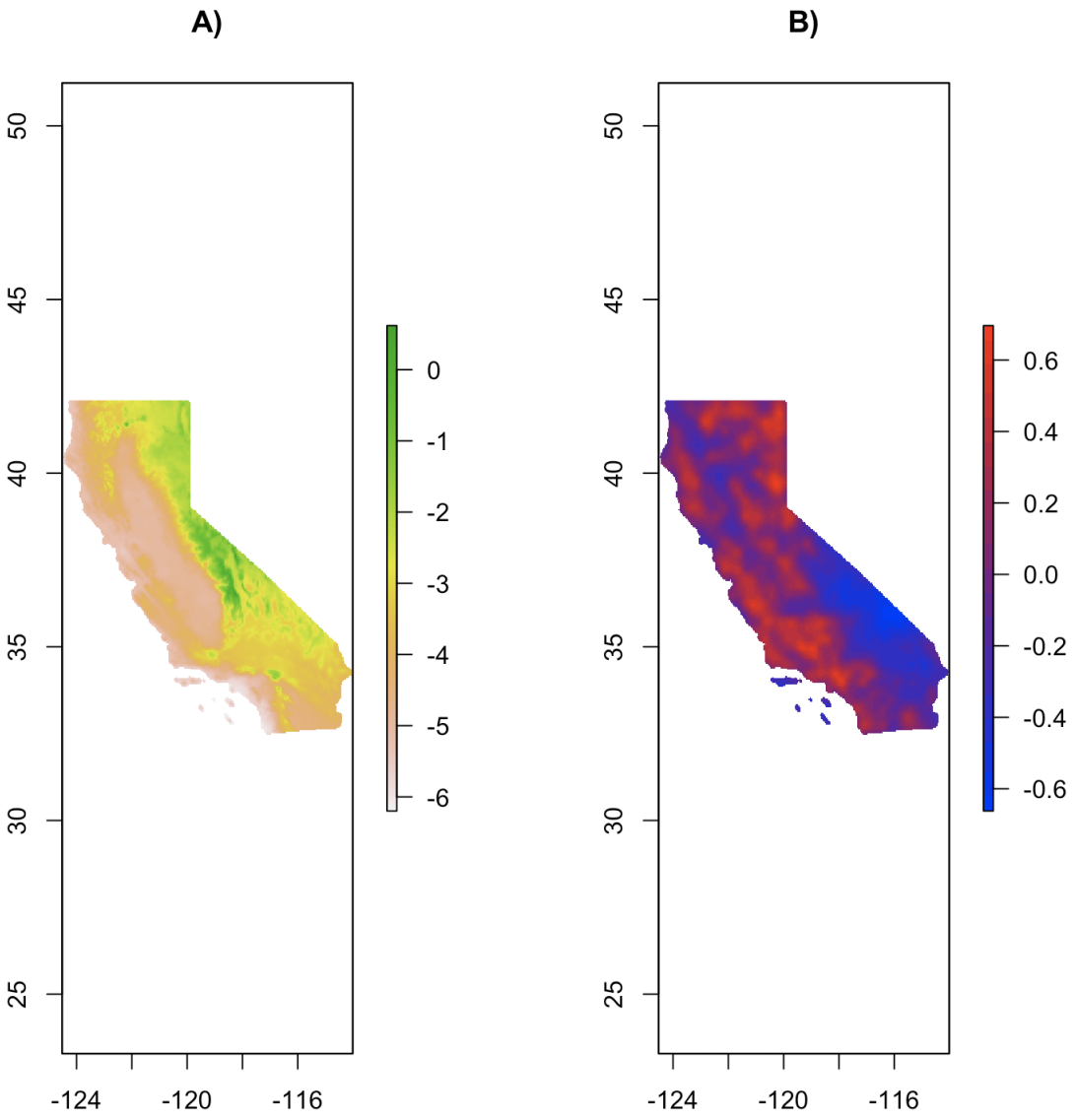

Figure 2.20: Comparison of A) covariate versus B) random effect influence on estimated log disease odds of plague in coyotes, showing A) the value contributed to the log odds by covariate effects, $z_\lambda(x)^T\beta_{+,\text{coyotes}} - z_\lambda(x)^T\beta_{-,\text{coyotes}}$ and B) that contributed by random effects, $(\alpha_{+,\text{coyotes}} \times w) - (\alpha_{-,\text{coyotes}} \times w)$.

Posterior variance in predicted risk for coyotes reaches a maximum value of 0.014, and shows a noticeable west-east gradient of increase, with variance sharply increasing towards the central and southeastern edge of the map (Figure 2.21). Note that, as one might expect, the variation in posterior variance roughly corresponds to variation in observation density. Elsewhere, however, the posterior variance remains low.

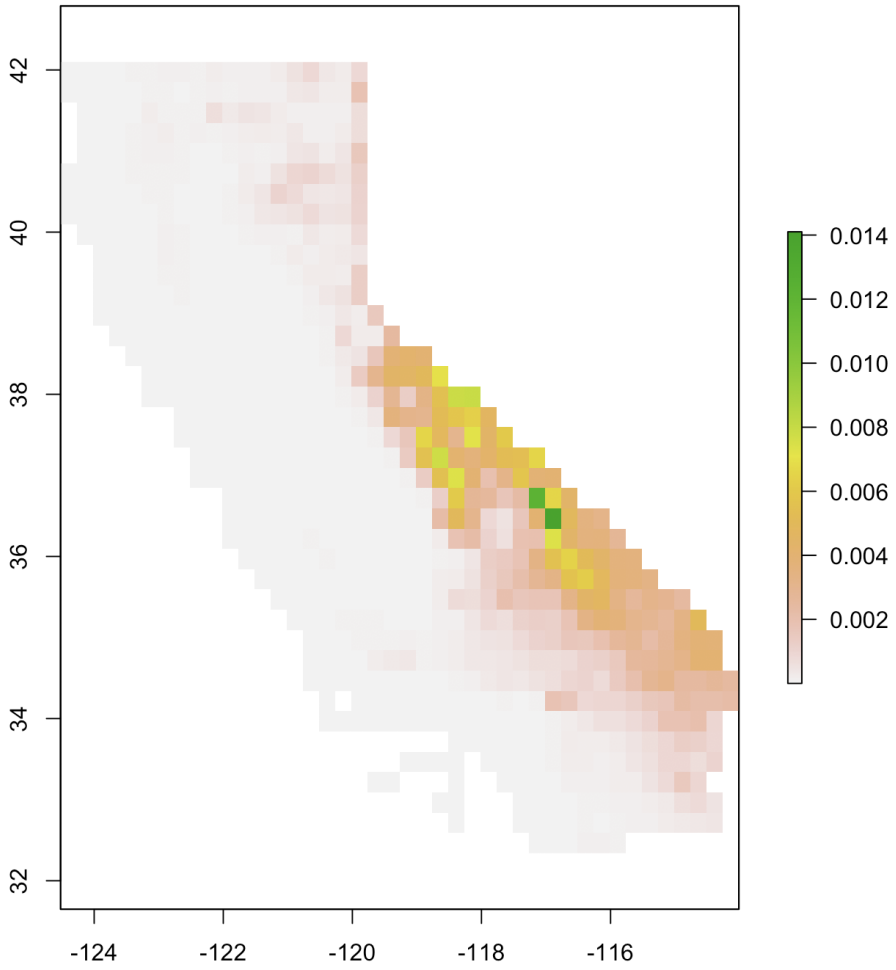

Figure 2.21: Posterior variance in predicted risk of plague in coyotes, as calculated by the multispecies model (2.1).

The risk map for plague in Sciurids obtained from model (2.1) ranges in value from 0.008 to over 0.101, which is, the probability that a Sciurid sampled at a particular

point will test positive for plague (2.22). Peak areas of risk fall along the Sierra Nevada mountain range and in the northeastern portion of the state. Additionally, neighborhoods of increased risk arise in Southern California, while the coastline shows a more subtly elevated level of risk.

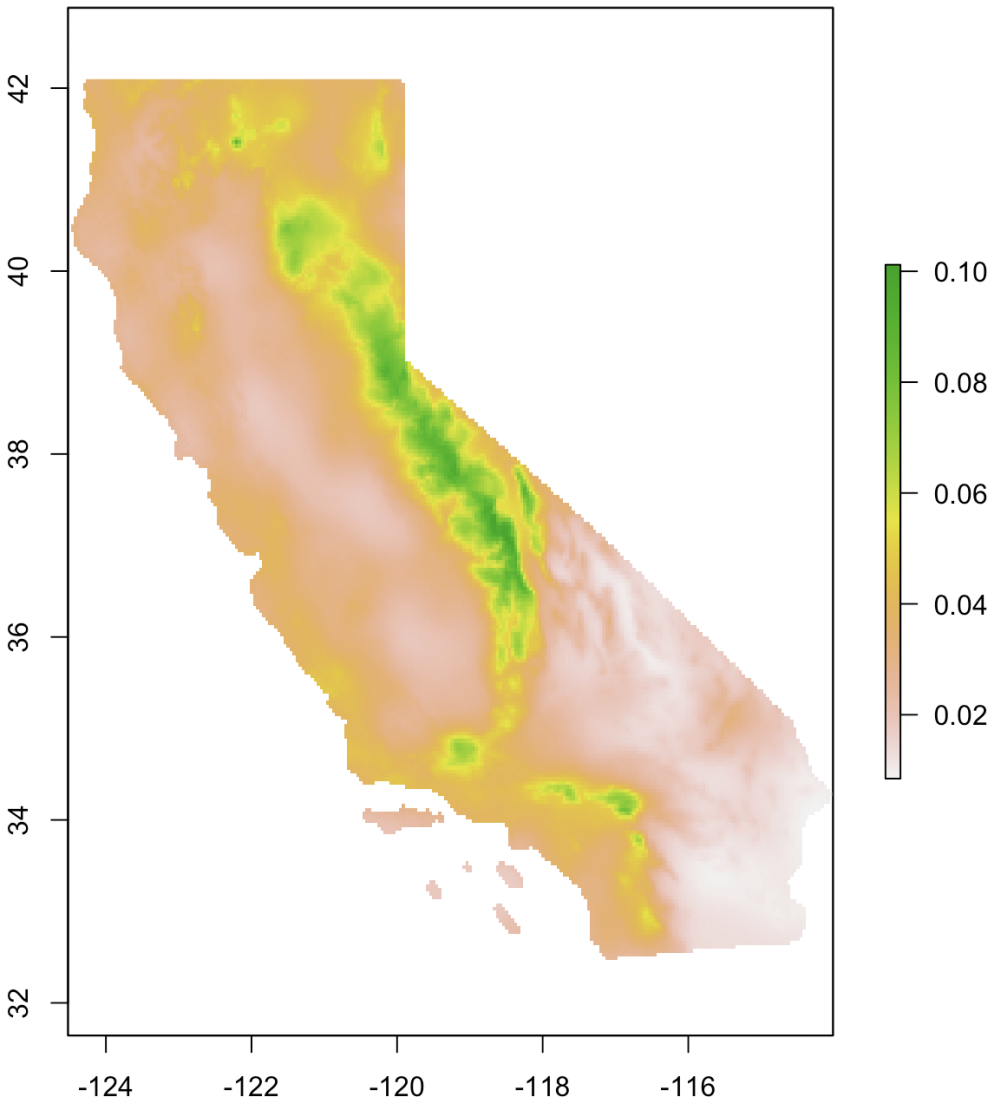

Figure 2.22: Sciurid risk map at 16 km^2 resolution obtained from the multivariate Gaussian process model.

We also consider an the distributions of observed cases and controls for Sciurids in relation to the estimated Sciurid risk map (Figure 2.23). It is apparent that the elevated band in risk running along the Sierra Nevada mountain range, between the 40th and 36th parallels, coincides with a strong presence of recovered cases.

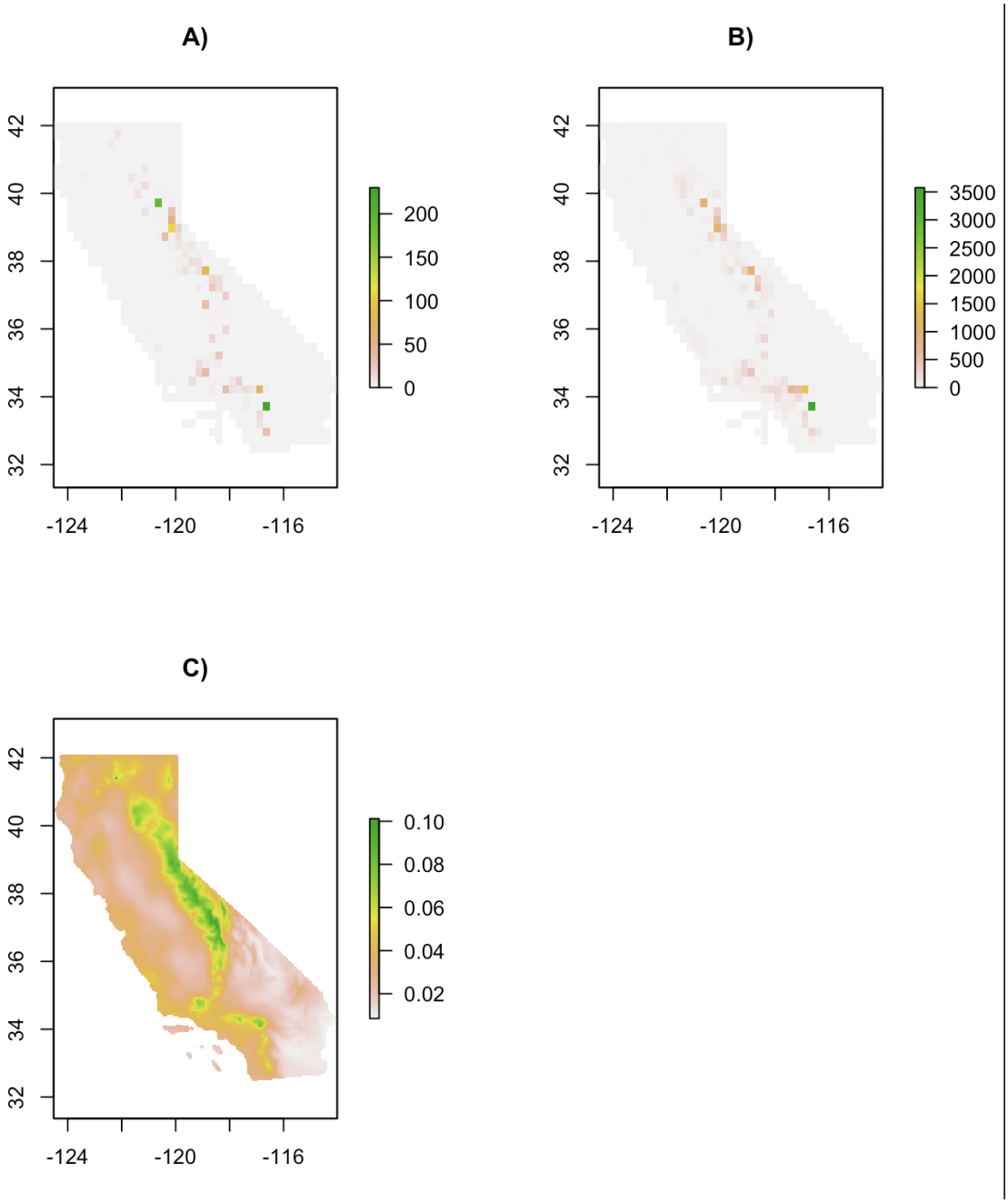

Figure 2.23: Distributions of A) case counts, B) control counts, and C) the estimated risk of plague in Sciurids between 1983 and 2015.

As we did for the coyote risk map, we next examine the relative contributions of covariates and random effects to the regions of high risk in the Sciurid plague map. From plotting rasters of the covariate contribution to log disease odds, $z_\lambda(x)^T\beta_{+,\text{sciurid}} - z_\lambda(x)^T\beta_{-,\text{sciurid}}$, next to that of the random effect contribution, $(\alpha_{+,\text{sciurid}} \times w) - (\alpha_{-,\text{sciurid}} \times w)$ we observe the high risk region along the Sierra Nevada mountains, between the 37th and 35th parallels, to be underpinned by both high covariate and random effect contributions.

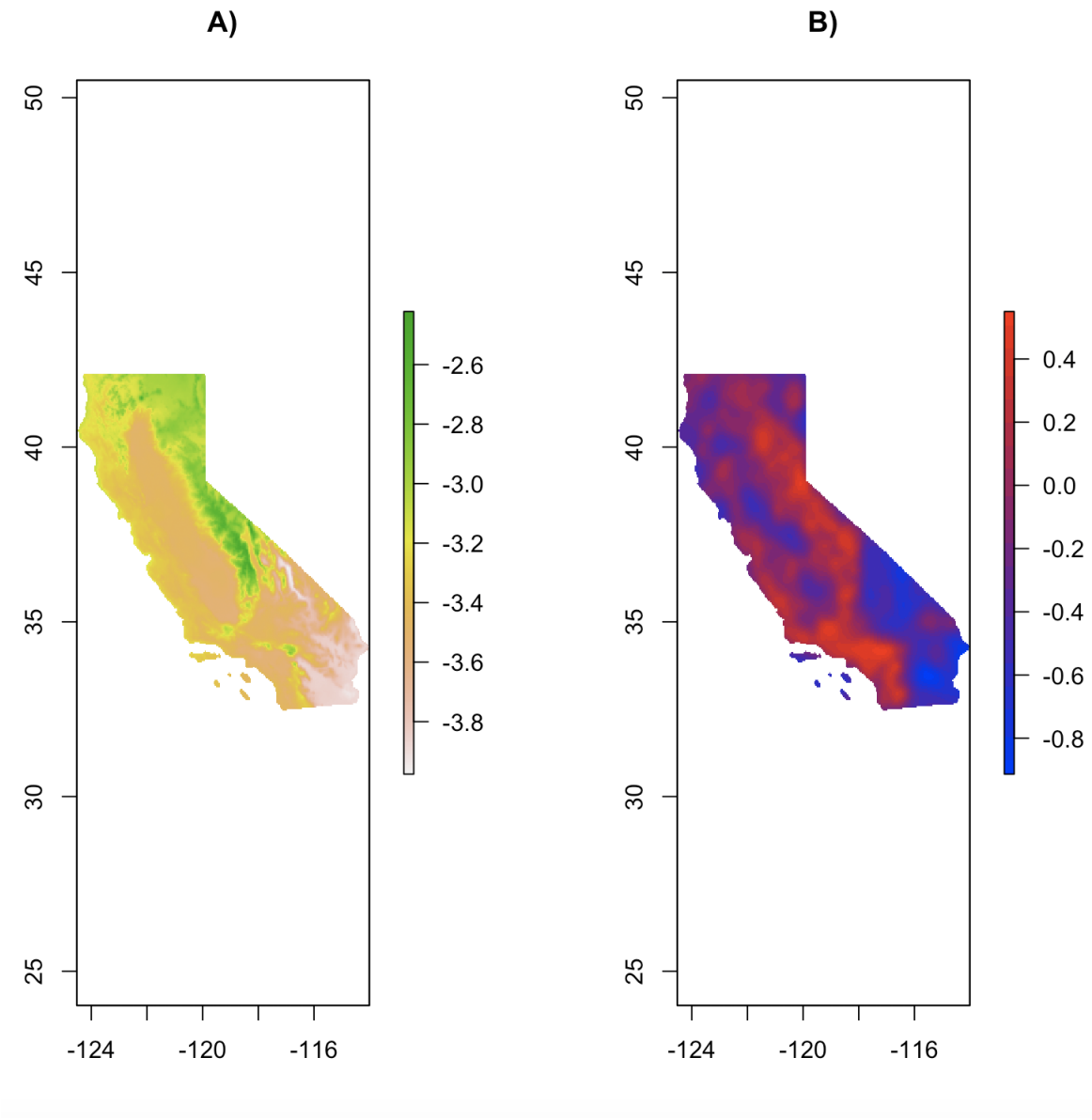

Figure 2.24: Comparison of A) covariate versus B) random effect influence on estimated log disease odds of plague in Sciurids, showing A) the value contributed to the log odds by covariate effects, $z_\lambda(x)^T\beta_{+,\text{sciurid}} - z_\lambda(x)^T\beta_{-,\text{sciurid}}$ and B) that contributed by random effects, $(\alpha_{+,\text{sciurid}} \times w) - (\alpha_{-,\text{sciurid}} \times w)$.

The posterior variance of predicted risk for plague in Sciurids remains low throughout the study region, not exceeding 1.176×10^{-4}, and is fairly homogeneously distributed across space (Figure 2.25), unlike that of coyotes which showed a clear east-west gradient.

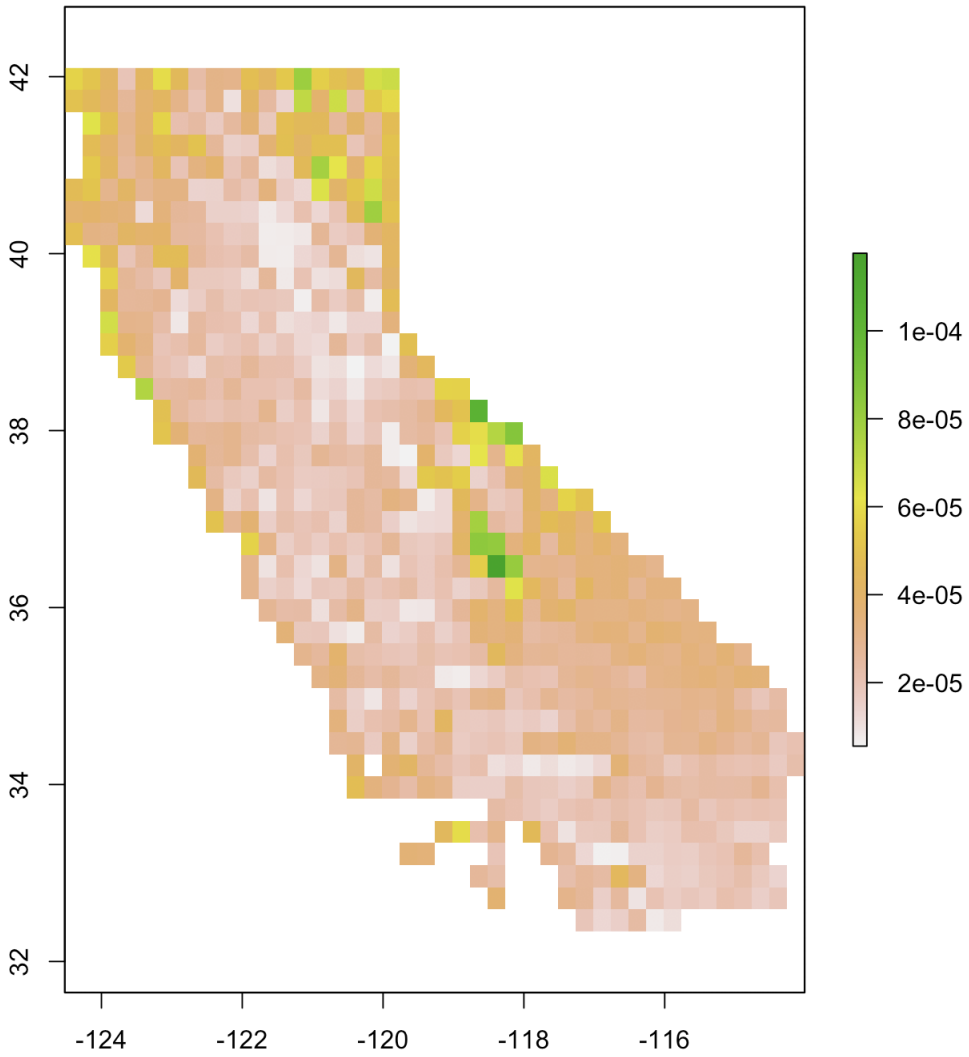

Figure 2.25: Posterior variance in predicted risk of plague in Sciurids, as calculated by the multispecies model (2.1).

Estimates of α_{ms}, the preferential sampling parameters in model (2.1), were found to be 0.634 and 0.576 for $\alpha_{+,\,\text{sciurid}}$ and $\alpha_{-,\,\text{sciurid}}$, respectively, while those for $\alpha_{+,\,\text{coyote}}$

and $\alpha_{-,\text{ coyote}}$ were 0.748 and 0.663, all of which had posterior variances at or below 0.002 (Table 2.21).

Parameter	Estimate	Posterior Variance
$\alpha_{+,\text{ sciurid}}$	0.634	< 0.001
$\alpha_{-,\text{ sciurid}}$	0.576	< 0.001
$\alpha_{+,\text{ coyote}}$	0.748	0.002
$\alpha_{-,\text{ coyote}}$	0.663	0.001

Table 2.21: Estimates of preferential sampling parameters for Sciurids and coyotes.

Next, in Table (2.22) we consider estimates of the T matrix from model (2.1), which we recall is the covariance matrix for the values of the spatial process w from all species groups at any given point in space. Under this interpretation, the off-diagonal element of T can be seen as providing the covariance in random effect values between Sciurids and coyotes, and as such, is one indication of the amount of information that can be shared between the two species groupings. This off-diagonal element is estimated as 7.878 with posterior variance 2.715. The first diagonal element, T (1,1) or the variance of the random effects for Sciurids, is estimated as 51.541 (with posterior variance 54.242), while that for coyotes is estimated as 12.169 (posterior variance: 5.835).

Parameter	Estimate	Variance
T (1,1)	51.541	54.242
T (1,2)	7.878	2.715
T (2,2)	12.169	5.835

Table 2.22: Estimates of the cross-correlation T matrix from the multivariate Gaussian process model, which has separable covariance structure $H \otimes T$.

We now turn to the portion of the analysis examining the quantitative differences in predicted risk between the joint model (2.1) and the univariate preferential sampling model, our benchmark method introduced in Project 1. We fit this model to the disease surveillance data from Sciurids and coyotes separately, both at a resolution of 614 km^2, from which we subsequently downscale the predicted risk to a resolution of

16 km^2. We first summarize the per-cell differences in predicted disease log odds for each model. Specifically, we calculate the average per-cell difference as

$$N^{-1}\Sigma_{i=1}^{N}(\hat{l}_{i,mvgp} - \hat{l}_{i,sep})$$

where N is the number of grid cells in the study region and $\hat{l}_{i,mvgp}$ are the estimated log disease odds for a given species grouping in the ith grid cell, as estimated by the multivariate Gaussian process model (2.1), and $\hat{l}_{i,sep}$ is that as estimated by the single-species model. The multispecies model predicted slightly greater log disease odds for coyotes compared to the single species model, with the average per-cell difference between the two models being 0.007, with a maximum difference of 0.422. On the other hand, model (2.1) predicted slightly lower disease odds for rodents relative to the single species model (average difference: -0.063, greatest difference, -0.459). Visual inspection of the per-cell differences in predicted log odds (Figure 2.26) shows that for coyotes, the overestimation in log odds by model (2.1) relative to the univariate tends to occur more at the higher end of the spectrum in estimated odds, while for rodents, the underestimation by model (2.1) also tends to occur to a greater degree for lower values of log odds.

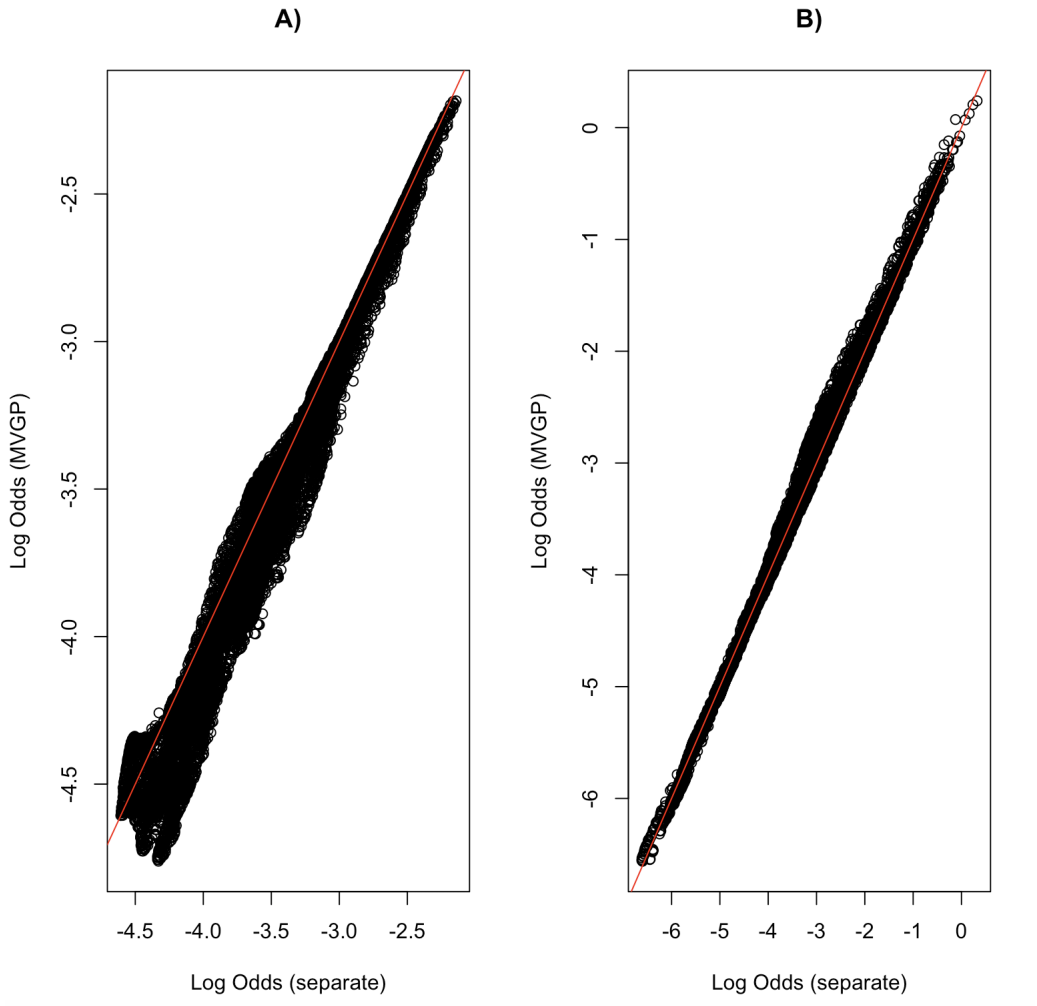

Figure 2.26: Comparison of predicted log disease odds for A) Sciurids and B) coyotes from 2 different models: a joint hierarchical model sharing information between the two species (MVGP) and separate analyses of each species (separate). The red diagonals are of slope 1 and intercept 0.

Lastly, the resulting differences in per-cell predicted risk between the multivariate and single species models are also visually summarized (Figure 2.27). Here, the average difference in predicted risk between model (2.1) and the univariate model was, for coyotes, 5.116×10^{-4}, with a maximum difference of 0.052, and, for rodents, on average -0.002, and at most 0.004.

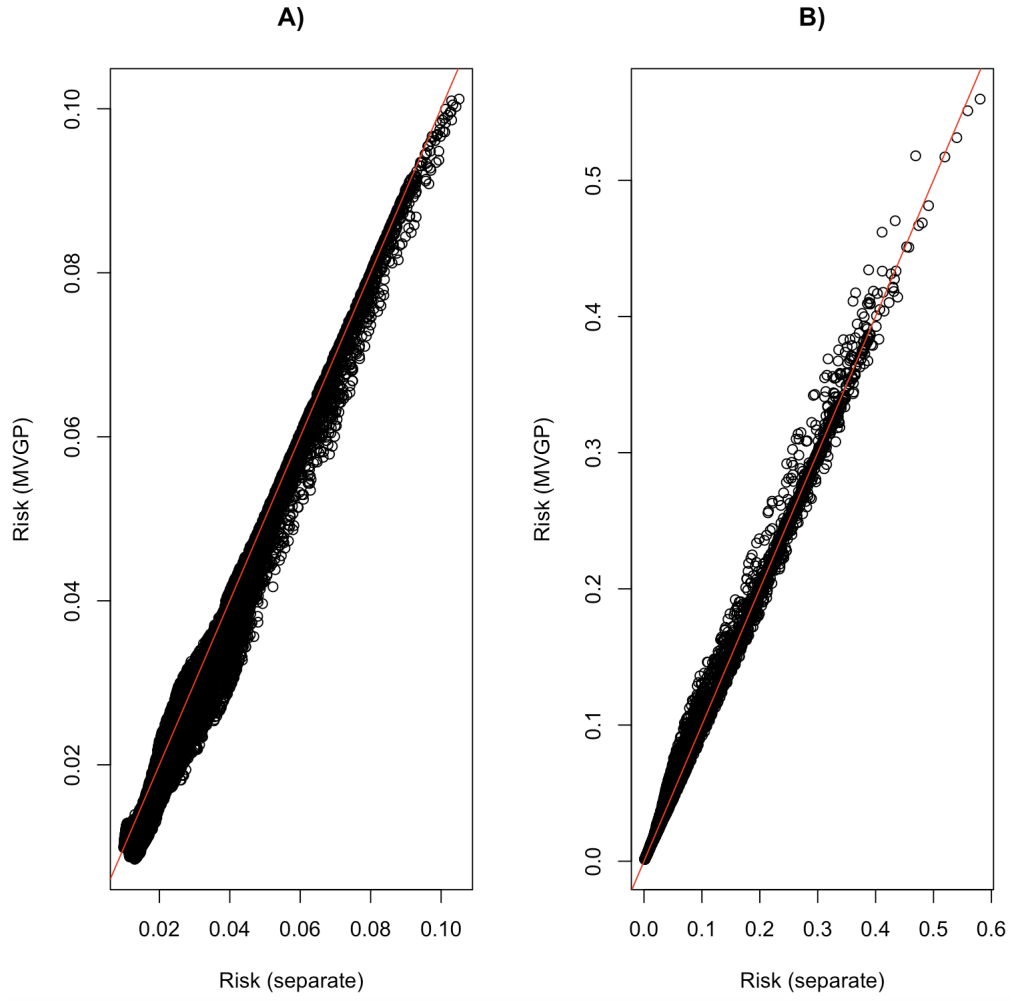

Figure 2.27: Comparison of predicted risk for A) Sciurids and B) coyotes from 2 different models: a joint hierarchical model sharing information between the two species (MVGP) and separate analyses of each species (separate). The red diagonals are of slope 1 and intercept 0.

2.5.4 Discussion

To put this analysis into perspective, we touch back upon Project 1, which applied a method to correct for preferential sampling in the plague surveillance data pertaining to Sciurids alone, over a 30 year window. This analysis extends our study of risk in the zoonotic hosts of plague by applying a newly developed, multispecies model jointly to both Sciurid and coyote surveillance data. The multivariate analysis not only updates the risk map for Sciurids, but also provides a risk map for coyotes from the same model. The intent of the multispecies model behind these risk maps is to both correct for the potentially biasing effects of preferential sampling on predicted risk, as well as to improve upon predictions made by the method from Project 1 through a strategy of sharing information between species in a hierarchical modeling framework. While bias in predicted risk cannot be assessed directly, since the true risk remains unknown, it is still helpful and possible to at least determine whether the two models produce quantitatively different estimates of risk. Thus, in this analysis we have compared models on the basis of the actual values of estimated risk. We have found that significant covariance exists between the latent spatial risk processes of each species, and that the two models produce slightly different estimates of risk.

We first turn our discussion towards the findings present in the estimated risk map for plague in coyotes. As mentioned above, this analysis marks the first instance in which we have produced a risk map for plague in this species, and as such, the general characteristics of this map are worthy of consideration. We seek to explain why the Sierra Nevada mountain range possesses a high level of predicted risk, despite the lack of samples recovered in this region. We argue that Figure (2.20), displaying the respective contributions of covariates $z_\lambda(x)^T\beta_{+,s} - z_\lambda(x)^T\beta_{-,s}$ and random effects $\alpha_{+,s} \times w_s(x) - \alpha_{-,s} \times w_s(x)$ to predicted log odds, shows that the climatic covariates are primarily responsible for this high predicted risk. In other words, this region holds

a climatic profile conducive to the presence of plague in coyotes, based on climate-risk associations observed in areas where samples were observed. The lack of recovered samples arises as a result of the nature in which coyotes are surveilled for plague. We recall that coyotes are often recovered as a result of livestock harassment responses or as roadkill. Naturally, the remote, mountainous Sierra Nevadas, with few roads or livestock, would not give rise to abundant samples in either scenario, despite the fact that coyotes do indeed inhabit this region. The precise nature of preferential sampling varies between Sciurids and coyotoes, as noted above, providing an interesting setting for our multispecies model.

Next consider the impact of preferential sampling as a force influencing the predicted risk map for plague in coyotes. We recall that preferential sampling refers to a stochastic dependence between the response being measured, which in this case refers to the abundances of cases and controls, and the locations of observation. In addition, the data would not be considered preferentially sampled if any association between response and location can be explained by covariates alone. For this last reason, it is valid to question whether these data do in fact show evidence of stochastic dependence between response and location. In the framework of model (2.1) this dependence arises through the sharing of the latent process w between the portion of the model describing case and control abundances and that describing the distribution of observation sites. If we take the response to be case or control abundances, then the response is independent from location if $\alpha_{+,s} = 0$ or $\alpha_{-,s} = 0$, respectively. As we see from Table (2.21) both are greater than zero with very low posterior variance. We interpret this to mean that the preferential sampling mechanism coincides with a preference for areas which tend to have greater abundances of not only cases, but also controls, for both Sciurids and coyotes. This finding is consistent with the expert knowledge of the wildlife biologists in CDPH.

Next, if we consider the response to be disease risk, rather than case or control counts alone, then preferential sampling would impact predicted risk when the random effect differential $\alpha_{+,s} \times w_s(x) - \alpha_{-,s} \times w_s(x)$ is nonzero. As the raster map of this difference in Figure (2.20) has shown, a substantial extent of the study region has positive values for $\alpha_{+,s} \times w_s(x) - \alpha_{-,s} \times w_s(x)$. Furthermore, as we have previously remarked, the elevation of predicted risk in the vicinity of Lake Tahoe seems to be driven primarily by this difference. Moreover, as is evidenced by Figure (2.20), the values of this difference can reach high positive values relative to the covariate differences. For these reasons we characterize the impact or strength of preferential sampling to be moderate to strong for coyotes. The practical consequences of this phenomenon would be realized if a model not accounting for preferential sampling were fit to the data, in which case predictions made at novel, unobserved locations would likely suffer inflated risk estimates, as was found in the analysis undertaken in Project 1.

The most prominent regions of increased risk fall along the Sierra Nevada mountain range and in the northeastern sector of the state. The side by side comparison of covariate versus random effect contributions to log disease odds in Figure (2.24) shows the increase of risk along the Sierra Nevada mountains, as well as in the northeastern corner, to be due to both covariates and random effects. The covariate contribution agrees with the existing knowledge of the CDPH for the climatic conditions necessary for plague, while the random effect contribution is consistent with the relative increase of sampling effort in this area compared to other portions of the state. With regard to the apparent presence of preferential sampling in the dataset, estimates of $\alpha_{+,s}$ and $\alpha_{-,s}$ for Sciurids are both positive, with low variance. We interpret this to mean that the sampling process tends to assign observation sites in areas which correspond to a higher overall abundances of Sciurids, which is consistent with the a priori belief of CDPH when defining its sampling plan. The contributions of preferential sampling to predicted log odds and risk are apparent in Figure (Figure 2.24B), showing the

differential $\alpha_{+,s} \times w_s(x) - \alpha_{-,s} \times w_s(x)$. We see widespread presence of moderate, positive values for this quantity, suggesting a moderate level of stochastic dependence between the response, risk, and spatial distribution of observation sites.

We now comment on the estimated values of the inter-species covariance matrix, T. We recall that because a separable model was specified for the multivariate Gaussian process in model (2.1), the covariance matrix for the random effects from both species at all points in the study is given by $H \otimes T$, where H is a matrix of spatial decay factors whose elements are computed by the correlation function, which was exponential for this analysis. The T matrix is interpreted as the covariance matrix of the random effect values for each species at any point in space. Consequently, the off diagonal element of T is interpreted as the interspecies covariance in random effects between rodents and Sciurids at any point in space. Nonzero positive covariance is apparent from this off diagonal element, estimated as 7.878 (posterior variance: 2.715). This estimate is of moderate magnitude relative the the marginal variance estimates for Sciurids and coyotes, and of much lower variance than either of those estimates. To place it in perspective, the off diagonal element is 15 and 65 percent of the variance estimates for Sciurids and coyotes, respectively. Thus, there is evidence for covariance between random effect values of the two species groupings, which we interpret ultimately as evidence for positive covariance in the risk for plague between Sciurids and coyotes. However, one notable a drawback of the separable approach is the assumption that the spatial correlation structure is the same for both species. Future work will examine adjustments to allow differing scales of spatial correlation within each species.

Having discussed the risk maps produced by the proposed model for each species grouping we now turn our attention toward the comparison of the results of this model with those estimated from the model developed in Project 1. Comparisons of

the per-cell predicted log odds between each model, visualized in Figure (2.26A), show a decrease in log odds predicted by the multivariate model relative to the univariate for a majority of cells. However, the multivariate model also predicts higher log odds for a smaller, yet nevertheless sizeable, number of cells. Turning toward the comparison of the estimated risks in Figure (2.27A), we see that risk estimates are generally quite close between the two models, with the exception of a number of points for which the values may differ by as much as over 57%. It is important to emphasize that this analysis alone does not shed light on which model has bias with regard to the true, unknown risk. But rather, we have established that there is some apparent difference in estimated log odds, small though it may be, between the single and multi-species models, and, for certain grid cells, an even greater difference in predicted risk. These differences provide fertile ground for further extensions of the models.

One prominent limitation of this study traces its origin to the notably different sampling methods for coyotes and Sciurids. We recall that whereas the recorded geocoordinates of Sciurids correspond precisely to the locations at which those specimen were trapped, the same does not hold for coyotes. These locational identifiers describing where coyotes were recovered vary in precision from high quality, such as latitude and longitude coordinates obtained from GPS, to verbal directions, such as the estimated mileage from a nearby road or other identifier. In the latter case, verbal directions were manually mapped to a set of latitude and longitude points. The potential for incorrect or misleading verbal description is a source of misclassification error which has been described and examined by Buller (2019), but merits further exploration as to how best to include such location uncertainty in our model-based framework. The posterior variances in estimated risk for plague in coyotes are likely lower than they should be, due to the failure to propagate this uncertainty through the levels of the hierarchical modeling framework. In addition to misclassification error, the analyses also fails to take into account location uncertainty, i.e., that the point at which a

coyote was recovered may not necessarily be near where it was exposed to plague, given the large ranges of these animals.

Lastly, and as noted briefly above, a key limitation of the proposed method lies in the separable covariance structure imposed upon the spatial random effects. As the second simulation of Project 2 has shown, model (2.1) is not always robust to failure of the separability assumption. To recap, when random effects were simulated from a linear coregionalization model rather than separable model, the resulting estimates of risk could be biased when the spatial decay factor and marginal variance of the coregionalization model were small, or in other words, a situation arose with very long range spatial interactions present. While the estimate of the spatial range in this analysis is small (posterior mean: 1.426, posterior variance: 0.053), the marginal variances of the rodent and coyote spatial processes are quite high, pointing to a situation in which a non-separable but stationary spatial process may still be modeled without catastrophic error. However, the assumption of stationarity remains a persistent threat, especially in a real world application, where it is likely that the associations between species may vary depending on the location in the study region, rather than mere separation between observations. Future work incorporating a more flexible, nonstationary covariance structure in the multivariate spatial process w may provide even better multispecies risk estimates.

Chapter 3

A Spatiotemporal Preferential Sampling Model

3.1 Introduction

Our first and second projects have focused on real world disease surveillance datasets whose sampling methods violate a key assumption of traditional geostatistical methods, namely that of independence between the response and the locations at which the response is measured. This phenomenon in which there is a stochastic relationship between where a process is measured and the value of that process is known as preferential sampling. When preferential sampling goes unaddressed in an analysis, negative consequences can strongly manifest themselves. An extensive literature (Diggle et al., 2010; Gelfand et al., 2012; Cecconi et al., 2016; Gelfand and Shirota, 2018), as well as the simulation chapters of Projects 1 and 2, have shown that preferentially sampling data can yield biased predictions for the response of interest. Intuitively this bias is not at all surprising, given that if one tends to favorably measure a process in

places where it is of high value, then extrapolating from such measurements may yield a positively biased set of predictions. While a rich body of work has been developed to address the problem of preferential sampling in relatively straightforward applications characterized by the measurement of a smooth surface over a finite set of points, such as air or soil pollution monitoring (Pati et al., 2011; Lee et al., 2011; Lee et al., 2015), few solutions (Cecconi et al., 2016; Rinaldi et al., 2015) are well equipped to address preferential sampling in a disease surveillance setting. The novelty of Projects 1 and 2 lies thus in extending the existing corrections for preferential sampling to the realm of disease surveillance, particularly applications concerning zoonotic diseases. However, the methods proposed in our first two projects are limited in one crucial respect. They do not account for temporal trends in the disease process or sampling mechanism.

Disease surveillance datasets often encompass recorded presences of cases (disease positive specimen) and controls (disease negative specimen) at particular points in space and time. For instance, the surveillance system operated by the California Department of Public Health (CDPH) targeting plague in the Sciurid family of squirrels, examined in Projects 1 and 2, consists of the locations and record dates of cases and controls over a 32-year window of time, between 1983 and 2015. However the methods proposed in the first two projects aggregate this data in time, resulting in a loss of temporal information which is problematic for several reasons. Firstly, over such a long count of years the risk of the underlying disease could shift under a variety of influences. In the case of zoonotic diseases, changes in both abiotic or biotic factors could impact both the distribution of the host species (e.g., Thorson et al., 2015) or even the dynamics of the disease itself, thereby affecting an expansion or contraction of the spatial extent of the disease. And secondarily, temporal fluctuations in sampling effort, as measured by the distribution of samples collected during some interval of time, could result in misleading conceptions of the confidence in our

estimates of risk. For instance, if the sampling effort decreases over time, resulting in fewer records being gathered of cases and controls over a smaller spatial extent, then the posterior variance of estimated risk for more recent years should be expected to increase, especially in areas where samples have not been conducted. If however the data are aggregated in time then regions which have not been sampled thoroughly for up to a matter of decades could still borrow information from the past that would, misleadingly, be taken as directly relevant to the present time. Indeed, just such an attenuation in sampling effort over time is apparent in the CDPH plague surveillance dataset (Figure 3.1). The importance of capturing temporal trends has spurred extensive development of spatiotemporal models of disease surveillance over the years (e.g., Waller et al., 1997; Quick et al., 2018). However, the existing body of work is still largely inadequate to address temporal effects in surveillance datasets that have been preferentially sampled. On the other hand, while several solutions exist to correct for preferential sampling in a time-agnostic context, there is a dearth of spatiotemporal preferential sampling methods. In short, methods exist separately to accommodate both preferential sampling, as well as spatiotemporal disease processes, but have not been combined. Our final project seeks to bridge this gap.

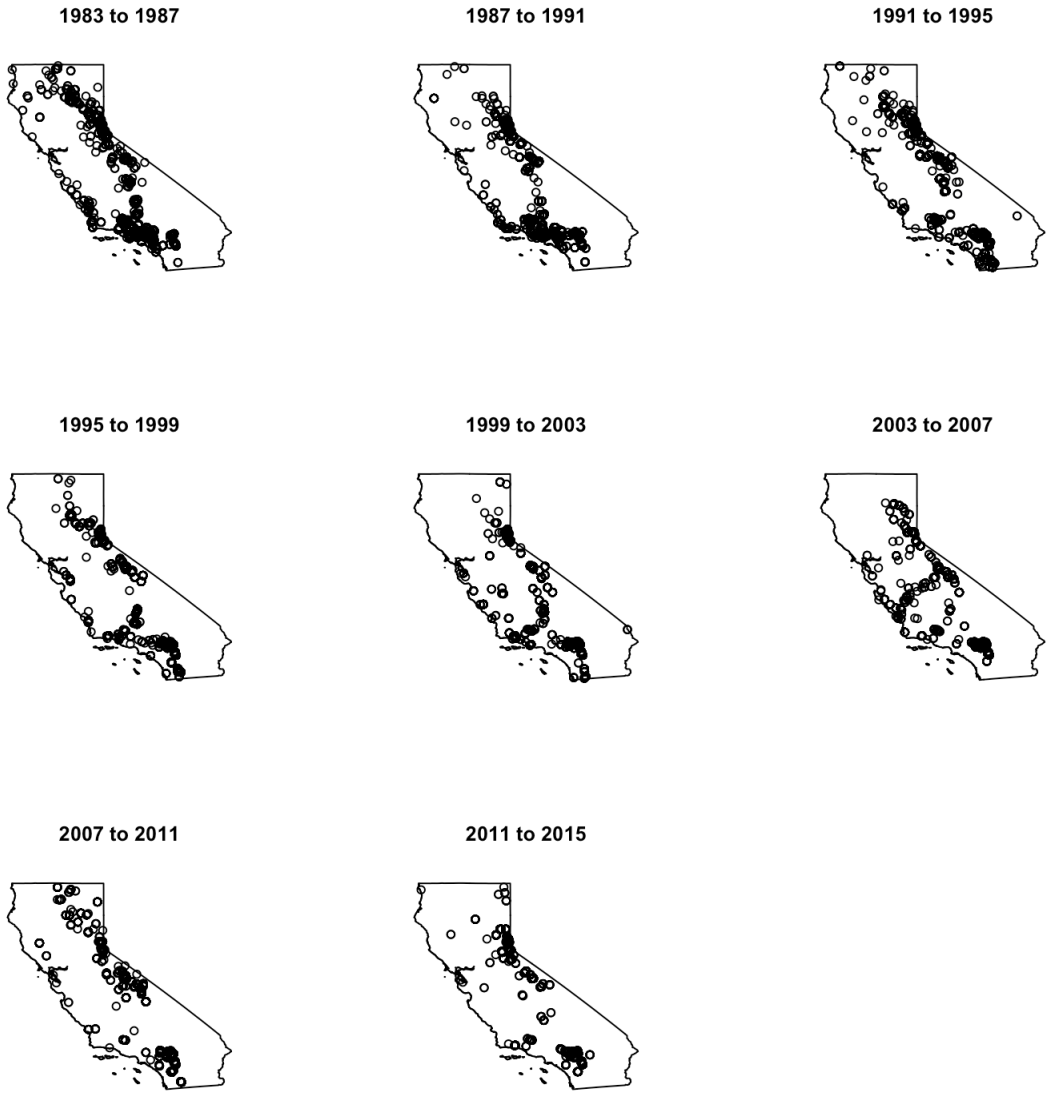

Figure 3.1: Distribution of sampling sites over time from the CDPH surveillance dataset targeting plague in Sciurids between 1983 and 2015.

The primary contribution of Project 3 is a discrete time spatiotemporal preferential sampling model. This new method enhances the model proposed in Project 1 to acommodate temporal changes in the underlying disease risk and sampling process. A secondary contribution of this project is the development of what we refer to as confidence maps, depicting our degree of certainty that the disease risk exceeds some particular threshold value for each pixel of space within the study region to a high degree of spatial resolution. Here, our measure of confidence is obtained from the posterior distribution of disease risk. For each high resolution pixel we calculate the posterior probability that risk exceeds the threshold value. Calculation of these values efficiently at high resolution required much experimentation with different spatial interpolation methods, which we detail in the Methods section of this project. The utility of these maps is immediately realized. In the previous two projects we had not combined our presentation of the posterior means of estimated risk with estimated posterior variances into a single figure. Instead, we separately reported estimated posterior means and variances. Consequently, it was impossible to directly formulate statements regarding the statistical significance of observed areas of high or low risk. The significance maps introduced here resolve this issue, offering clearly interpretable visualizations regarding the posterior distributions of estimated risk. Furthermore, an added practical benefit of these maps is their utility for informing sampling design, particularly by suggesting areas which would be valuable to sample in the future. These areas are identified to be of potentially elevated risk for disease, but lacking in sufficient samples to draw conclusions with confidence.

This project also includes an extensive simulation study probing the performance of the proposed model, as well as comparing it to the time-aggregated preferential sampling model developed in Project 1. In this study we demonstrate the clear superiority of the proposed method in terms of error in estimated log disease odds when the data are simulated under increasing, decreasing, and alternating temporal trends

in disease risk and sampling effort. We follow these simulations with an analysis section applying the proposed spatiotemporal model to the Sciurid plague surveillance dataset maintained by CDPH. This plague dataset provides fertile grounds for application of the proposed model due to its preferentially sampled and temporally referenced nature. In the analysis, we uncover important temporal trends in the underlying disease risk and sampling process that are overlooked by the time-aggregated approach.

3.2 Methods

3.2.1 Introduction

Throughout this dissertation one of our core objectives has been to model real world disease surveillance systems as realistically as possible, first by accounting for preferential sampling in the data collection process and then by sharing statistical information between ecologically related species in a multivariate modeling framework. However, when a surveillance system covers an extensive temporal range, it may be deleterious to simply aggregate the data in time, as we have done until this point. Aside from failing to capture fluctuations in the dynamics of the disease of interest, or changes in the sampling process by which data are generated, temporal aggregation becomes especially problematic in the surveillance of zoonotic diseases due to the fact that species distributions can also change over time in response to changing biotic and abiotic factors (Thomas and Lennon, 1999; Brommer, 2004; Maclean et al., 2008). By oversimplifying reality, neglecting the temporal structure of the surveillance data may pose a threat to the quality of statistical estimates and predictions.

Our primary methodological objective of this third project is to develop a spatiotem-

poral model for preferentially sampled disease surveillance data, with a motivating dataset originating from zoonotic disease surveillance in particular. A key ultimate product of our proposed method is a series of temporally referenced disease risk maps. Our secondary methodological objective is to introduce what we term significance maps, which incorporate measures of statistical uncertainty into our presentation of disease risk, in particular by visually displaying our confidence that disease risk exceeds a certain pre-determined threshold value over space and time. These developments are intended to be limited to applications containing disease surveillance data pertaining to a single species or species grouping (i.e. collection of species assumed functionally or ecologically equivalent), with the understanding that they could be extended in the future to a joint analysis of disease risk across multiple species along the methodological lines presented in Project 2.

We assume our readers are by this time acquainted with preferential sampling, which we briefly review here. Preferential sampling is a data collection strategy in which observation sites or samples of some process are deliberately assigned to areas thought to be of high value for that process. Applications characterized by preferential sampling often include air pollution monitoring (Lee et al., 2011), mineral exploration (Veneziano and Kitanidis, 1982), species distribution modeling (Gelfand and Shirota, 2019), disease surveillance (Cecconi et al., 2016; Rinaldi et al., 2015), and real estate pricing (Paci et al., 2019). In many cases, the practical consequences of preferential sampling manifest themselves in biased spatial predictions (Diggle et al., 2010; Pati et al., 2011; Lee et al., 2011; Gelfand et al., 2012; Lee et al., 2015), at times considerably so, when traditional geostatistical models assuming independence between locations and the response are employed. To mitigate this bias a number of approaches have been proposed to correct for preferential sampling, the majority of which revolve around a joint modeling framework in which the distribution of observation sites and response are stochastically related by way of a common latent spatial process (Dig-

gle et al., 2010). Project 1 modified the basic idea of this framework to encompass disease surveillance data. Project 2 proposed a multivariate extension of Project 1 in which disease surveillance data from multiple different species were combined in a joint hierarchical model. In Project 3, we now return to the shared latent process framework of the first project, only now introducing the ability to capture temporal trends in disease risk and the sampling mechanism.

The spatiotemporal preferential sampling model proposed in this project relies on a discrete-time, shared latent process framework. In this model a latent spatiotemporal process describes both the distribution of observation sites over space and time as well as the abundances of cases and controls at each sampled location. Temporal structure is captured through fixed effects for time intervals, which may be of arbitrary width, but were taken to be 5-year windows in our analysis chapter. The temporal nature of the latent process is intended to model changes in both the underlying disease risk and sampling process over time. But before providing a more formal description of this model, we first review the existing solutions for spatiotemporal modeling in disease surveillance as well as in a closely related domain, namely species distribution modeling, before showing how this existing literature stops short of spatiotemporal preferential sampling.

3.2.2 Spatiotemporal Modeling in Disease Surveillance

An abundance of statistical methods have been developed to identify and characterize spatiotemporal trends in disease surveillance data. Here we review the most prominent of these methods, namely scan statistics, generalized linear mixed models, process models, disease mapping models, and point process models.

Scan Statistics

The first class of methods we consider is that of the scan statistics, which primarily serves to identify areas in space or time that suggest disease clustering, or significant elevations in the incidence of cases. Originally developed to identify anomalies of case incidence purely from a temporal perspective (Naus 1965), the concept of the scan statistic has been extended to incorporate spatial (Openshaw et al., 1987) and spatiotemporal (Kulldorff, 2001) data with a straightforward but powerful approach. For instance, the Kulldorf and Nagarwalla (1995) spatial scan statistic identifies clusters of cases by imposing circles of differing radii over the study region to demarcate areas of differential disease rates. For each possible circle centroid, circles of increasing radii up to some pre-determined limit are formed. For each of these generated circles Z, case counts inside and outside the perimeter of Z are summed, which are assumed under the alternate hypothesis to follow Poisson distributions whose rates differ between the inside and outside of the circle. Under this assumption of differential rates the likelihood of the data $L(Z)$ may be found, prompting the calculation of a likelihood ratio

$$S = \frac{\max_Z\{L(Z)\}}{L_0}$$

which is taken to be the scan statistic, where L_0 is the likelihood of the data under the null hypothesis, that is, no clustering or equal rates throughout space. P values for S may be subsequently obtained via Monte Carlo hypothesis testing. The spatial scan statistic S readily offers inspiration for a spatiotemporal scan statsitic (Kulldorff, 2001), which imposes cylinders, rather than circles, over the study region in such a way that the vertical height of the cylinder represents its duration in time.

While scan statistics are integral to the exploration of spatial or spatiotemporal disease surveillance datasets (Reis et al., 2007), they are are not particularly well suited to address preferentially sampled data, as traditionally they make no correction for biased observation processes, but rather analyze the data as observed. In addition, scan statistics typically do not include covariate information, which may considerably influence the distribution of cases. A more covariate driven approach is offered by the next class of spatiotemporal disease surveillance methods, generalized linear mixed models.

Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) naturally lend themselves to the analysis of disease surveillance datasets, being both relatively straightforward and simple to implement while offering the ability to account for important covariate effects, in contrast to scan statistics. In disease surveillance settings generalized linear mixed models typically regress case and, possibly, control counts on covariates along with temporal or spatial effects. Here, temporal trends are commonly modeled with fixed effects, such as day of the week, or month, or year, depending on the temporal scale of the disease of interest, while space is often encoded by random effects for the areal unit or region in which an observation falls (e.g., Kleinman et al., 2004; Bradley et al., 2005; Johnson 2008). Of note here is the fact the spatial random effects in these models are, in fact, not spatially structured in the sense that their correlation is oftern not a function of distance but rather based on neighborhood structures between administrative regions. To more directly model correlation as a function of geographic distance, we turn to the spatiotemporal process models.

Spatiotemporal Process Models

In Project 1 we examined the spatial process model, which describes a spatially smooth, normally distributed response $Y(s)$ at location $s \in \mathbb{R}^2$ as the sum of a deterministic mean function $\mu(s)$ and residual component $w(s) + \epsilon(s)$, where $w(s)$ and $\epsilon(s)$ are spatially structured and unstructured residuals, respectively:

$$Y(s) = \mu(s) + w(s) + \epsilon(s)$$

The deterministic mean $\mu(s)$ is typically a function of spatial covariates, $\mu(s) = x(s)^T\beta$, and the $w(s)$ follow a Gaussian process with mean 0 and stationary covariance function $k(s, s'; \sigma^2, \theta)$, which calculates the covariance between points s and s' as a function of distance, of the marginal variance σ^2, and of the range parameter θ, which controls the rate of decrease of covariance as distance increases. The intent for $w(s)$ is to capture additional spatial variation not explained by the mean component $\mu(s)$. Residuals $\epsilon(s)$, are assumed independent with mean zero and variance τ^2, referred to as the *nugget* effect. The nonspatial residuals $\epsilon(s)$ are often interpreted as measurement error or noise accompanying repeat measurements at a particular location, or as micro-scale variability, i.e., variation in the response at distances smaller than the distance between sites observed in the data.

In the following sections we examine spatiotemporal extensions of this modeling structure. But first we note that while disease surveillance data typically consist of case and control counts over a series of observation sites, rather than a normally distributed response over a discrete set of points, the spatiotemporal process model above may be easily modified to acomodate this new distributional assumption. For instance, the spatial structure of the residuals, $w(s)$, which is our true concern here, may be

incorporated into the structural component of a Poisson regression model as

$$Y(s) \sim \text{Poisson}(\lambda(s))$$
$$\log(\lambda(s)) = x(s)^T\beta + w(s)$$

In the following sections we describe the spatiotemporal process models in terms of a normally distributed response, with the understanding that the models discussed could easily be translated into a distributional form more common to disease surveillance applications.

Discrete Time Models

For point-referenced, spatially continuous data whose response is assumed to be normally distributed, a general space-time model is described by Bannerjee, Carlin and Gelfand (2014) as

$$Y(s,t) = \mu(s,t) + e(s,t)$$

where $Y(s,t)$ is the response at location s and time t, $\mu(s,t)$ is the mean at s and t, and $e(s,t)$ is the residual. Often $\mu(s,t)$ is chosen to be a linear function (perhaps through a link function) of spatial and temporal covariates such as $\mu(s,t) = x(s,t)^T\beta$, for covariate vector $x(s,t)$ and parameters β, which may be spatially varying, temporally varying, or both. In the most general case $e(s,t)$ is written as $w(s,t) + \epsilon(s,t)$, where $w(s,t)$ is a spatiotemporal process, with mean zero, and $\epsilon(s,t)$ is a Gaussian white noise process (Bannerjee et al., 2014). Here, $\epsilon(s,t)$ can be thought of as a temporal

extension of the "nuggett" effect of the spatial process model.

A number of different space-time models can be obtained according to the specification of the residual process $e(s,t)$. A key distinction to make in spacetime modeling is whether time is indexed continuously or discretely. In the former case, measurements are recorded at potentially any point in time, whereas discrete indices provide measurements only within pre-defined windows of time, such as hours, days, or years. In the discrete time case, Gelfand et al. (2003) suggest three different formulations of the residual process.

In the first, $e(s,t) = \alpha(t) + w(s) + \epsilon(s,t)$, where $\alpha(t)$ and $w(s)$ are temporal and spatial effects, respectively, and $\epsilon(s,t) \sim N(0, \sigma^2_\epsilon)$ are independent white noise error terms. If time were discretized into intervals $t = 1, 2, \ldots, T$, then a number of different modeling options present themselves. The parameters $\alpha(1), \ldots, \alpha(T)$ could simply be fixed effects, as in Bailey et al. (1963) or Knight et al (1995). Alternately, an autoregressive specification would model $\alpha(t+1) = \rho\alpha(t) + \eta(t)$, with independent error term $\eta(t)$ distributed as $\eta(t) \sim N(0, \sigma^2_\alpha)$.

The second approach utilizes independent time series at each location. Here, $e(s,t) = \alpha_s(t) + \epsilon(s,t)$, where $\alpha_s(t)$ are temporal effects nested within each site, and $\epsilon(s,t)$ again arise from a Gaussian white noise process. As in the first approach, $\alpha_s(t)$ can be assigned an autoregressive structure, $\alpha_s(t+1) = \alpha_s(t) + \eta_s(t)$, where $\eta_s(t)$ are white noise error terms. In the third approach, $e(s,t) = w_t(s) + \epsilon(s,t)$, where $w_t(s)$ are spatial random effects nested in time, that is, independent spatial processes for each time interval.

One salient drawback of discrete time models lies in the pitfalls surrounding the choice of the width of the time discretization interval. If time is discretized to an overly wide or narrow extent then the true scale of the spatial process may not align with that of

the time intervals, and consequently, model predictions and inference may suffer. We now turn to an alternative modeling approach which avoids the necessity of specifying a time discretization window.

Continuous Time Models

In continuous time models we assume the residuals $e(s,t)$ to arise from a continuous spatiotemporal process, rather than the sum of independent spatial or temporal components as in the discrete time case. The challenge in this context becomes how to specify $e(s,t)$ such that its covariance function is valid. That is, for any finite set of locations and points in time, $\mathcal{S}$, the covariance matrix of the random effects realized from $e(s,t)$ over $\mathcal{S}$ must be positive definite (Bannerjee et al., 2014). Note that for this purpose we cannot simply choose a valid covariance function on $\mathbb{R}^3$, as may initially seem an intuitive option, due to the difference in scales on which space and time operate. Instead a valid covariance function is typically reached through a combination of spatial and temporal covariance functions.

To obtain a spatiotemporal covariance function that is valid, i.e., symmetric and positive definite, as well as stationary, depending only on the separation between two points in space and time, rather than their absolute locations in space or time, a common (e.g., Mardia and Goodall, 1993) approach is the separable model:

$$\text{Cov}(e(s,t), e(s',t')) = \sigma^2 \rho^{(1)}(s - s'; \phi)\rho^{(2)}(t - t'; \psi)$$

where $e(s,t)$ and $e(s',t')$ are values of the spatiotemporal process at locations s, s' and times t, t', and $\rho^{(1)}$ and $\rho^{(2)}$ are valid two and one dimensional correlation functions, respectively, which model the spatial and temporal aspects of covariance. In this

construction the covariance between two points decreases as either their distance in space or separation in time increase, as modulated by the range parameters ϕ and ψ, which control the scale of the spatial and temporal correlation.

While the assumption of separability is appealing from a computational standpoint, entailing less burden than other continuous-time alternatives, it may be overly restrictive for certain spatiotemporal processes. In such cases where greater flexibility is desired, a nonseparable covariance structure can be employed. For instance, if $e(s,t) = e_1(s,t) + e_2(s,t)$, where e_1 and e_2 are independent spatiotemporal processes with separable spatiotemporal covariance functions, then the covariance function for $e(s,t)$ is nonseparable (Bannerjee et al., 2014).

However the greater flexibility of these continuous time models, both separable and nonseparable, comes at the cost of a greater computational burden than what is required by discrete time models. For instance, if a dataset consists of measurements at n points in space each of which have s repeated observations in time, then the covariance matrix of a separable, continuous time spatiotemporal process is of dimension $(n \times s)^2$, in contrast to a discrete time model $e(s,t) = \alpha_t + w(s)$ with fixed effects α_t, which entails a covariance matrix merely of dimension $n \times n$. Avoiding this computational cost will inform the design of our proposed model.

Disease Mapping Models

Yet another approach to capture spatial and temporal trends in disease surveillance data in a more sophisticated fashion than that offered by GLMMs comes in the form of disease mapping models, specifically, the conditional autoregressive (CAR) framework. CAR models were developed with respect to areal data, consisting of measurements gathered over areal units, such as counties or states, in contrast to the

point-referenced measurements associated with spatial process models. The seminal CAR application to disease surveillance data was provided by Besag, York and Mollie in a seminal paper (1991). In this model, the random effect for the kth areal unit in the study region, denoted ψ_k, is taken as the sum of a spatially structured random effect, ϕ_k, and a spatially unstructured random effect θ_k:

$$\begin{aligned}
&\psi_k = \phi_k + \theta_k \\
&\phi_k | \phi_{-k}, W, \tau^2 \sim N\Big(\frac{\Sigma_{i=1}^K w_{ki}\phi_i}{\Sigma_{i=1}^K w_{ki}}, \frac{\tau^2}{\Sigma_{i=1}^K w_{ki}}\Big) \\
&\theta_k \sim N(0, \sigma^2)
\end{aligned}$$

where the spatially structured random effects ϕ_k are conditionally dependent on their neighbors, whose spatial relationships are encoded in the weight matrix W, with the typical convention that the kith element w_{ki} of W is equal to 1 if the areal units i and k are adjacent. Several alternative versions of this famous "convolution" or "CAR-BYM" model have been developed. For instance, Leroux et al. (2000) proposed a CAR model eliminating the random effects θ_k and incorporating a parameter ρ controlling the strength of spatial dependence. Other popular variants of the spatial CAR model include Lee and Mitchell (2012) and Lee and Sarran (2015). Especially germane to our discussion of spatiotemporal disease surveillance are the space-time CAR models proposed by Waller et al. (1997) and Knorr-Held and Besag (1998), both applied to rates of lung cancer in the state of Ohio. Multivariate extensions of this space-time model, featuring the ability to jointly model multiple responses over time, have also been employed Knorr-Held (2000).

Point Process Models

The last methodology for characterizing spatial and temporal trends which we will consider is the point process framework, which describes the distribution of random collections of points. Point processes have been featured extensively in spatiotemporal disease surveillance studies (Brix and Diggle, 2001; Diggle, 2006; Kottas et al., 2009; Robertson et al., 2010; Li and Guan, 2014), many of which trace their origin to the spatiotemporal Cox process (Cox 1955). For instance, Diggle et al. (2005) develop a spatiotemporal model for the analysis of data derived from surveillance of gastrointestinal infections, as recorded by calls to National Health Service. Letting X denote the point pattern of cases and $R(x,t)$ a latent spatiotemporal process, the authors propose the point pattern of cases to be, conditional on the value of the latent process, and inhomogeneous point process with intensity function $\lambda(x,t)$

$$X \sim \mathcal{IPP}(\lambda(x,t))$$

The spatiotemporal intensity function $\lambda(x,t)$ is decomposed into spatial, temporal, and spatiotemporal components

$$\lambda(x,t) = \lambda_0(x)\mu_0(t)R(x,t)$$

Here $\lambda_0(x)$ is a smooth, nonparametric, spatially varying surface while $\mu_0(t)$ is a parametric function capturing temporal variation, incorporating day of the week and seasonal fixed effects, and $R(x,t)$ is a spatiotemporal stochastic process intended to describe local variation in the presentation of cases, which is modeled as a stationary log Gaussian Cox process:

$$\log(R(x,t)) = S(x,t)$$
$$S(x,t) = \mathcal{GP}(-0.5\sigma^2, \rho(u,v))$$

where $S(x,t)$ is a Gaussian process with mean $-0.5\sigma^2$, variance σ^2, and correlation function $\rho(u,v) = \text{Corr}(S(x,t), S(x-u,t-v))$, which the authors assume to have a separable structure, i.e., $\rho(u,v) = \rho_x(u)\rho_t(v)$, where ρ_x and ρ_t are valid spatial and temporal correlation functions, respectively. With this approach, the authors can identify the normal pattern of spatial and temporal variation in presentation of cases, and subsequently locate hotspots in space or time constituting departures from normality, specifically by evaluating the predictive probability $\Pr(R(x,t) > c|\text{data until time } t)$, for some pre-defined threshold c. In our proposed method we adopt elements of both spatiotemporal point processes and spatiotemporal process models. But first, we round out our review of spatiotemporal modeling approaches with a discussion of species distribution models, which in many ways shed light on our modeling efforts with regard to zoonotic diseases.

3.2.3 Spatiotemporal Species Distribution Modeling

Modeling efforts regarding the surveillance of zoonotic diseases may gain insight from the domain of species distribution models, since both applications are characterized by many of the same complexities of data collection, such as opportunistic sampling or preferential sampling. To begin, numerous species distribution models have been put forward to characterize species distributions from a time-agnostic standpoint, including both parameteric methods, typically generalized linear models (Royle et al., 2007; Latimer et al., 2006; Maggini et al., 2006) and non-parametric methods,

such as the maximum entropy (Maxent) species distribution model (Phillips et al., 2006). However increasing attention has been devoted to temporal modeling, which is particularly important for understanding how species will respond to future changes in habitat, climate, or other perturbations (Walther et al., 2002). Many of the spatiotemporal models applied to species distributions show parallels to those of disease surveillance. For instance, a common approach is the use of generalized linear mixed models with fixed effects for units of time and random effects for area of the study region (e.g., Link and Sauer, 2007; Helser et al., 2004). In fisheries stock assessment, the abundance and distribution of acquatic life is commonly analyzed with a two step GLMM (Pennington, 1983), which first models the probability of a nonzero catch, and subsequently models the abundance or density of nonzero catches (Thorson and Ward, 2013; Ward et al., 2015; Helser et al., 2004). When spatial and temporal terms are included as fixed or random effects then spatiotemporal interactions have been modeled by including interaction terms, such as region $\times$ year (Thorson and Ward, 2013). In addition to fixed or random effects, temporal trends have also been addressed with time series models, such as in (Ward et al., 2015), who utilize spatiotemporal effects $\epsilon_t(s) \sim \text{Normal}(\rho\epsilon_{t-1}(s), \Sigma)$ with a temporally autoregressive mean and spatially structured covariance matrix Σ, or Hooten and Wikle (2005), who feature a vector autoregressive process. Aside from these parametric modeling approaches, also worthy of note are the nonparamteric, machine learning based models developed for species distributions, such as the STEM framework (i.e. spatiotemporal exploratory model) of Fink et al. (2010), which trains an ensemble of base models (such as decision trees) on restricted spatiotemporal extents, and forms the predicted value of the species distribution as an aggregation of the outputs of the base model.

Lastly, while recent efforts have addressed preferential sampling in species distribution modeling (Conn et al., 2017; Pennino et al., 2019), these have not done so in a spatiotemporal context. In general, methods to address preferential sampling in

spatiotemporal analyses are scarce, with the few existing efforts avoiding the joint modeling framework of Diggle et al. (2010) from which the majority of spatial preferential sampling models are derived. For instance, when analyzing preferentially sampled air pollution monitoring data over time, Yu et al. (2015) correct for preferential sampling by incorporating secondary information into a knowledge synthesis framework (Bayesian Maximum Entropy), where the secondary information describes pollutant levels in unsampled regions. While this approach does adjust for preferential sampling over time the secondary information on which it relies may not always be obtainable. Shaddick and Zidek (2014) examine evidence for preferential sampling in an air pollution monitoring network over a 30-year period. The monitoring network of witnessed drop in the number of sites over time as pollution levels fell, while sites in more polluted areas tended to be retained. By grouping sites according to length of operation and fitting different spatiotemporal process models to each group, the authors showed that sites which were retained longer had higher pollutant levels. However, the focus of their work was to provide evidence that preferential sampling was present, rather than detailing a specific modeling solution to confront it. Thus, the area of spatiotemporal preferential sampling remains largely unexplored, which motivates our modeling efforts here.

3.2.4 Proposed Method

Our proposed method is intended to address a preferentially sampled disease surveillance dataset $\mathcal{D}$ of the form $\mathcal{D} = (\mathcal{X}_+, \mathcal{X}_-)$, where $\mathcal{X}_+$ is a set consisting of triples each of which provides the latitude, longitude, and date of observation of a disease positive specimen (case). That is, $x = (\text{latitude}, \text{longitude}, \text{date}) \forall x \in \mathcal{X}_+$. Likewise, $\mathcal{X}_-$ is a set identically defined except with respect to disease negative specimen (controls). For the purposes of a general modeling definition, the temporal resolution of the dates

of observation may be specified at any scale that is not of finer resolution than the discretization of time chosen for model fitting. To discretize time, we suppose that the duration of the study, defined as the difference between the maximum and minimum observation dates in $\mathcal{D}$, is partitioned into T equally sized, non-overlapping windows, which we index by the variable $t = 1, \ldots, T$.

Our modeling intent now becomes how to describe the distribution of cases and controls within each time interval t. Let $\mathcal{X}_{t+}$ and $\mathcal{X}_{t-}$ denote the observed point patterns of cases and controls in the study region during the tth time interval, respectively. We assume that $\mathcal{X}_{tm}$, where $m \in \{+, -\}$ denotes the mark or disease status of points, follow log Gaussian Cox processes with intensity functions $\lambda_{tm}(x)$:

$$\mathcal{X}_{tm} \sim \mathcal{LGCP}(\lambda_{tm}(x))$$

Here, we have adopted the notational convention of using the term (x) to identify that the function $\lambda_{tm}(x)$ varies over space, and we have taken $x \in \mathbb{R}^2$ to represent a point in two dimensional space. We model the intensity functions $\lambda_{tm}(x)$ to be log-linear in spatially varying fixed effects $z_\lambda(x, t)$ along with a spatiotemporal latent process $w(x, t)$:

$$\log(\lambda_{tm}(x)) = z_\lambda(x, t)^T \beta_m + \alpha_m \times w(x, t)$$

Here we have included the index t in our notation for the fixed effects $z_\lambda(x, t)$ to emphasize that $z_\lambda(x, t)$ may differ among different time intervals. While a number of options are available for the specification of the spatiotemporal process $w(x, t)$, here we choose a simple additive model consisting of a fixed effect u_t specific to each time interval t, and a spatiotemporal, stationary Gaussian process $w(x)$:

$$w(x,t) = u_t + w(x)$$

The log-intensity function can thus be written as $\log(\lambda_{tm}(x)) = z_\lambda(x,t)^T\beta_m + \alpha_m \times (u_t + w(x))$. We note that more sophisticated forms for the spatiotemporal process could be specified, such as the use of an autoregressive time series $u(t)$ to describe the temporal trend, rather than fixed effect covariates:

$$\begin{aligned} w(x,t) &= u(t) + w(x) \\ u(t+1) &= \rho u(t) + \eta(t) \\ \eta(t) &\sim N(0, \sigma_\eta^2) \\ w(x) &\sim \mathcal{GP}(0, k(.,.;\theta,\phi)) \end{aligned}$$

Our choice of the simpler spatiotemporal process, $w(x,t) = u_t + w(x)$, was motivated by the fact that the dataset considered in our analysis chapter suggested a discretization of $T = 7$ time intervals, a number arguably too small for an autoregressive structure to be profitable. Thus, we proceed with the simpler model $u_t + w(x)$, with the understanding that more complex spatiotemporal processes can be considered in future efforts.

But rather than directly working with the likelihood of the point processes $\mathcal{X}_{tm}$, which would involve a cumbersome numerical approximation of the integrals of the intensity functions $\lambda_{tm}(x)$, we approximate the point processes as Poisson random variables over a discretization of the study region. We suppose that the study region is discretized into K equally sized grid cells, $g_1, \ldots, g_K$, of sufficiently small area such that the intensity functions $\lambda_{tm}(x)$ are constant over all g_k. Consequently, from the

definition of log Gaussian point processes, Y_{kmt}, the number of cases ($m = +$) or controls ($m = -$) in any grid g_k for a particular time interval t is distributed as

$$\begin{aligned} Y_{kmt}|w(x_k) &\sim \text{Poisson}(\int_{g_k} \lambda_{mt}(s,t)ds) \\ &\approx \text{Poisson}(|g_k|\lambda_{mt}(x_k,t)) \end{aligned}$$

To correct for preferential sampling in the shared latent process framework of Diggle et al. (2010) we must now specify the distribution of observation sites in terms of the spatiotemporal process $w(x,t)$. For this purpose we define indicator variables $\kappa_{xt} \in \{0,1\}$, representing whether the grid cell with centroid at point x is observed during the tth time interval. We model the indicators as Bernoulli random variables with probability of success $\xi(x,t)$:

$$\begin{aligned} \kappa_{xt}|\xi(x,t) &\sim \text{Bernoulli}(\xi(x,t)) \\ \text{logit}(\xi(x,t)) &= u_t + w(x) \end{aligned}$$

The inclusion of $w(x,t)$ in both the components of the model describing the distribution of observation sites and the distributions of observed cases and controls is intended to correct for the effect of preferential sampling. We thus summarize the fitted model by combining our components describing the distributions of observations, cases, and controls as follows:

$$
\begin{aligned}
&\kappa_{xt}|\xi(x,t) \sim \text{Bernoulli}(\xi(x,t)) \\
&\text{logit}(\xi(x,t)) = u_t + w(x) \\
&w(x) \sim \mathcal{GP}(0, k(.,.;\theta,\phi)) \\
&Y_{xmt}|w(x) \sim \text{Poisson}(\lambda_{mt}(x,t)) \\
&\log(\lambda_m(x,t)) = z_\lambda(x,t)^T\beta_m + \alpha_m \times (w(x) + u_t)
\end{aligned}
\tag{3.1}
$$

where Y_{xmt} denotes the count of cases or controls ($m \in \{0,1\}$) over the tth time interval in the grid cell with centroid at point x, $z_\lambda(x,t)^T$ are fixed effect covariates, $w(x)$ is a stationary, mean-zero Gaussian process with covariance function k parametrized by range θ and marginal variance ϕ, u_t is a fixed effect parameter specific to the tth time interval, and κ_{xt} are indicator variables representing whether the grid cell whose centroid lies at point x is observed during interval t.

Estimation

Model (3.1) is fit in the Bayesian framework by a Markov Chain Monte Carlo solution consisting of separate Hamiltonian Monte Carlo samplers for $\alpha_+, \alpha_-, \beta_+, \beta_-, u_t$ and spatial random effects w, along with a Metropolis-Hastings random walk update for the spatial range θ, and finally with a Gibbs sampling update for the marginal variance ϕ. The samplers involved have been custom implemented in the *R* statistical programming language, version 3.4.3, without making use of pre-built MCMC packages due to the unique nature of the model being fit.

To complete our Bayesian specification of model (3.1) we provide the prior distributions for its parameters. We assign normal priors to $u_t, \alpha_+, \alpha_-, \beta_+$ and β_-. In the

analyses and simulations conducted here we assign uninformative priors for u_t, β_+ and β_-, with large prior variances, as is often the case when estimating slope parameters in Bayesian analysis. For the spatial range parameter θ of the exponential covariance function, due to the constraint of $\theta > 0$, the chosen proposal distribution was the log-normal distribution, which has density function

$$q(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big(- \frac{(\log(x) - \mu)^2}{2\sigma^2} \Big)$$

for $x > 0$. The mean of the proposal distribution was taken to be the log of the current value of θ. That is, given the gth MCMC sample $\theta^{(g)}$, the proposed next value in the Markov chain was distributed as

$$\theta^{(g+1)} \sim \text{Log-Normal}(\log(\theta^{(g)}), \sigma^2)$$

The proposal standard deviation σ^2 was manually tuned to yield acceptance rates close to 0.5. The acceptance probability was calculated as

$$\min\Big(1, \frac{\ell(w^{(g)}; \theta^{(g+1)})}{\ell(w^{(g)}; \theta^{(g)})} \times \frac{p(\theta^{(g+1)})}{p(\theta^{(g)})} \times \frac{q(\theta^{(g)})}{q(\theta^{(g+1)})} \Big)$$

where $\ell(w^{(g)}; \dots)$ is the log likelihood of the spatial random effects w at the gth MCMC iteration, and $p(\dots)$ is the prior density of θ, taken here to be the gamma distribution, and $\frac{q(\theta^{(g)})}{q(\theta^{(g+1)})}$ is the ratio of log-normal densities from the current and proposed values of θ, which would have cancelled out had the proposal distribution been symmetric.

We assign an Inverse-Gamma prior distribution to the spatial marginal variance ϕ, in

order to make use of the fact that the conditional distribution of ϕ given the random effects w is also Inverse-Gamma. Specifically,

$$\phi|w \sim 1/\text{Gamma}(N/2 + a, w^T H^{-1} w + b)$$

where N is the number of elements of w, a and b are the shape and scale parameters of the prior distribution of ϕ, and H is the correlation matrix of the random effects w. ϕ can thus be updated at each step of the Markov chain by drawing a sample from $\phi|w$, rather than relying on a more computationally expensive Metropolis-Hastings or Hamiltonian Monte Carlo sampler.

Unlike typical spatial process models assuming normally distributed responses, model (3.1) places a Poisson distribution upon the observed responses of case and control counts, thereby precluding a closed form conditional distribution for the spatial random effects w upon which Gibbs sampling would rely. Consequently, the random effect vector w is updated via a Hamiltonian Monte Carlo (HMC) sampler. This technique was chosen given its effectiveness for updating high dimensional, spatially structured parameters. Briefly, HMC proposes new parameter states by simulating the dynamics of Hamiltonian physics, which describe the total energy of a system as the sum of potential and kinetic energies. Here, the potential energy of the system is taken as the negative log likelihood of the current parameter state, and a new parameter state is reached by evaluating the gradient of the potential energy. Thus, by essentially incorporating information from the gradient of the log likelihood in its proposal step, HMC is able to explore the parameter space more efficiently and, crucially, account for spatial correlation between elements of the parameter vector, a distinction which would not hold if a Metropolis-Hastings random walk updating strategy were employed here. Hamiltonian Monte Carlo algorithms are parametrized by a

step size parameter, related to the degree of change undertaken in the proposal step, and length parameter, which controls the number of iterations for which Hamiltonian dynamics are simulated in each proposal. The step size parameter was automatically tuned by the strategy of dual averaging, presented by Hoffman and Gelman (2010), which alters the step size after each proposal step based on comparing the current acceptance probability with the desired acceptance rate. The length parameter was manually tuned.

Hamiltonian Monte Carlo samplers were also assigned to separately update α_+, α_-, β_+, β_-, and u_t $(t = 1, \ldots, T)$. Despite the fact that these parameters are low dimensional and spatially uncorrelated, HMC sampling showed less inter-sample correlation than Metropolis-Hastings random walk and so was ultimately preferred. A more detailed technical description of HMC and the technique of dual-averaging can be found in the methods section of Project 1.

3.2.5 Significance Maps

The last ambition of our final project seeks to provide a visual display of our confidence in identifying above average areas of disease risk from the estimates of model (3.1). For this purpose we wish to calculate the posterior probability that the risk of any point $x \in \mathbb{R}^2$ of the study region exceeds some predetermined threshold c:, given the observed data $\mathcal{D}$

$$p(x|\mathcal{D}) = \Pr(r(x) > c|\mathcal{D})$$

The intent is to display values of $p(x|\mathcal{D})$ calculated at the centroids of each grid cell in a high resolution raster of the study region. For simplicity, we color code each grid

cell according to three levels of confidence, indicating whether the above posterior probability exceeds 0.95, 0.50, and 0.25, with the understanding that these numbers are easily configurable. Such a map would identify areas where were are strongly, moderately, or weakly confident that disease risk exceeds c. The remaining portion of this section details the calculation of the posterior probability $p(x|\mathcal{D})$ at unobserved points in space.

We recall that model (3.1) contains realizations of a spatial Gaussian process $w(x)$ at each centroid of the discretization of the study region. Consequently, if the study area is discretized into K grid cells then the covariance matrix of $w(x)$ is of dimension $K \times K$. The computational burden associated with inverting this matrix several times throughout the MCMC routing involved in fitting model (3.1) precludes an overly fine discretization. However, a high resolution map of posterior probabilities is still desired, which presents an obvious challenge. For instance, a raster of the state of California at $4km^2$ resolution corresponds to 25,701 grid cells, or a spatial covariance matrix with 660,541,401 elements. Project 1 addressed this challenge by interpolating the posterior mean of the random effects realized from $w(x)$. However, since this procedure was merely applied to the posterior mean of $w(x)$, but not each MCMC sample, it did not yield posterior samples for each high resolution point in the study region, and thus, cannot be used to calculate the desired probability $p(x|\mathcal{D})$.

To overcome the shortcomings of the downscaling method from Project 1, we propose an alternative algorithm to obtain posterior risk samples at high resolution. The basic idea of this algorithm is to interpolate each posterior sample of $w(x)$ over all high resolution points in the study region, after which posterior samples of risk may be generated. For each sample, fast interpolation was performed through nonparametric regression with kernel smoothing, using estimated values of $w(x)$ at observed points as the response and the latitudes and longitudes corresponding to those points as pre-

dictors. The specific nonparametric regression technique used here is that described by Li and Racine (2004), which relies on a generalized product kernel that takes the product of univariate kernels applied to each explanatory variable, i.e. latitude and longitude. The univariate kernel chosen here is a second order Gaussian, given by

$$k(z) = \exp(-z^2/2)/\sqrt{2\pi}$$

for $z = (x_i - x)/h$ with $h > 0$. The bandwidth h was obtained through least squares cross validation. The R package npreg was used to fit and cross validate the nonparametric regression model. Our proposed algorithm to efficiently generate posterior samples is summarized as follows:

Data: Posterior samples $(w^1, \ldots, w^G)$, with $w^g = (w_1^g, \ldots, w_K^g) \forall g \in \{1, \ldots, G\}$
Posterior samples of all other model parameters $\Psi = (\psi^1, \ldots, \psi^G)$,
Geo-coordinates of observed locations $\mathcal{C} = ((x_1, y_1), \ldots, (x_K, y_K))$,
Geo-coordinates of unobserved locations $\mathcal{C}^* = ((x_1^*, y_1^*), \ldots, (x_L^*, y_L^*))$
Result: array$[g \times L]$
for $g \in (1, \ldots, G)$ **do**
    Train model regressing w^g on $\mathcal{C}$: $w_i^g = f(x_i, y_i)$;
    Predict w at unobserved locations: $\hat{w}_i^{*g} = f(x_i^*, y_i^*)$;
    Calculate posterior risk sample at unobserved locations: $\hat{r}_i^{*g} = \text{risk}(\hat{w}_i^{*g}, \psi^g)$;
    Store sample: array[g,] $= (r_1^{*g}, \ldots, r_L^{*g})$;
end

Algorithm 1: Generation of posterior predicted samples

Since this algorithm re-interpolates values of the spatial random effects for each posterior sample, it captures the most crucial sources of uncertainty, namely those associated with random effects w and the parameters governing w (marginal variance and range). The algorithm does omit uncertainty associated with the interpolation

step, $\hat{w}_i^{*g} = f(x_i^*, y_i^*)$, but this uncertainty is small relative to that of w and so we ignore it for the time being.

The end product of this algorithm is series of posterior samples of risk for points corresponding to the centroids of a high resolution discretization of the study region. The posterior probability $p(x|\mathcal{D}) = \Pr(r(x) > c|\mathcal{D})$ is estimated as the fraction of posterior samples associated with point x which are above the threshold c:

$$\hat{p}(x|\mathcal{D}) = G^{-1} \sum_{g=1}^{G} I(\hat{r}^g(x) > c)$$

where $\hat{r}^g(x)$ is the gth posterior sample of risk associated with point x. As an example of the practical utility of the above algorithm, when implemented in parallel on an 8 core Macbook Pro, roughly 30 minutes were required to interpolate 10,000 MCMC samples over the state of California at $4km^2$ resolution.

3.3 Simulations

3.3.1 Introduction

This simulation study assesses performance of the proposed spatiotemporal preferential sampling model in comparison to benchmark methods which either do not account for the sampling process giving rise to the data, or do not account for temporal trends related to disease risk and the sampling process. To garner a broader view of the comparative performances of these models under different scenarios, we evaluate them over simulated datasets encompassing a range of spatiotemporal trends. The trends we consider are increases, decreases, and alternating fluctuations in the

disease risk surface over time. To emphasize performance with respect to the temporal nature of the data, evaluation of the models is based upon root mean squared error in the predicted log disease odds over each time interval of the simulated data. In addition to probing comparative model performance, this simulation study also seeks to observe characteristics, such as failure, of convergence in the proposed spatiotemporal model. The impetus for this aim stems from Project 1, which showed the tendency for correlation to arise between the latent process of the preferential sampling model and other parameter values. In the spatiotemporal model, with the introduction of new temporal parameters, it is reasonable to suspect the possibility of even greater parameter correlation, or even failure of convergence, which we wish to diagnose.

Models Compared

We compare our proposed spatiotemporal preferential sampling model against two benchmarks, which either ignore the temporal nature of the data or the preferential sampling process. The first of these is the preferential sampling method developed in Project 1. This model is similar in structure to (3.1), only stripped of its temporal aspects. To fit this model, counts Y_{xmt} and observation indicators κ_{xt} are aggregated across time intervals as $Y_{xm} = \Sigma_{t=1}^{T} Y_{xmt}$ and $\kappa_x = I(\Sigma_{t=1}^{T} \kappa_{xt} > 0)$. This approach, referred to in the results as the "aggregated preferential sampling model" or "aggregated" model for short, is summarized as

$$\kappa_x|\xi(x) \sim \text{Bernoulli}(\xi(x))$$
$$\text{logit}(\xi(x)) = w(x)$$
$$w(x) \sim \mathcal{GP}(0, k(.,.;\theta,\phi))$$
$$Y_{xm}|w(x) \sim \text{Poisson}(\lambda_m(x))$$
$$\log(\lambda_m(x)) = z_\lambda(x)^T\beta_m + \alpha_m \times w(x)$$

The second reference model does not account for the preferential sampling process behind the data but does distinguish data within different time intervals. For a given disease status m (cases or controls) and within a given time interval t and grid cell x which has been observed by the disease surveillance system, this model describes observed abundances Y_{xmt} as Poisson random variables with rates $\lambda_{mt}(x)$, which are log-linear in fixed effect covariates $z_\lambda(x,t)$. We write this model as

$$Y_{xmt} \sim \text{Poisson}(\lambda_{mt}(x))$$
$$\lambda_{mt}(x) = \exp(z_\lambda(x,t)^T\beta_m)$$

That is, we fit separate Poisson regression models to the case and control counts at observed grid cells within each time interval. We refer to this strategy as the "temporal Poisson" approach.

Evaluation Metrics

Models are compared on the basis of root mean squared error in estimated log disease odds over each time interval. Log disease odds are derived from estimates of the case and control intensity functions, $\lambda_{+,t}$ and $\lambda_{-,t}$, in the following way. Suppose we let $r_t(x)$ be the probability an individual sampled at location x in time interval t is disease positive. Then the case intensity function is $\lambda_{+,t}(x) = r_t(x)p_t(x)$, where $p_t(x)$ is the (unknown) spatial population density of individuals in time t, while the control intensity is given by $\lambda_{-,t}(x) = [1 - r_t(x)]p_t(x)$. Consequently the disease odds are the ratio of intensity functions

$$r_t(x)/[1 - r_t(x)] = \lambda_{+,t}(x)/\lambda_{-,t}(x) \tag{3.2}$$

and so, log disease odds are given by $\log(\lambda_{+,t}(x)) - \log(\lambda_{-,t}(x))$. Estimated log disease odds are thus obtained from estimates of the temporal case and control intensity functions for each model. The spatiotemporal model (3.1) and temporal Poisson model both directly yield estimates for time and mark specific intensities λ_{mt}. When considering the aggregated preferential sampling model, which merely estimates overall intensities λ_m ($m \in \{+, -\}$), we assume $\lambda_m = \lambda_{mt}$ for all t, after which we may proceed with estimating the temporal risk surfaces according to (3.2).

Lastly, to calculate the temporal root mean squared error, for a study region discretized into K grid cells, letting $o_t(x_k)$ denote the true disease odds at time t for grid cell x_k, and $\hat{o}_t(x)$ denote the estimated disease odds, we calculate the root mean squared error in estimate log odds over the interval as

$$\sqrt{K^{-1}\Sigma_{k=1}^{K}(o_t(x_k) - \hat{o}_t(x_k))^2}$$

Thus, each time interval t will be associated with an RMSE metric.

3.3.2 Data

For this study a total of 75 datasets with differing temporal trends were simulated from model (3.1). 25 of these datasets were simulated each under and increasing temporal trend, 25 under a decreasing trend, and 25 under an alternating trend. We recall from model (3.1) that this latent process is given by $u_t + w(x)$, where u_t is a parameter specific to time interval t and $w(x)$ is a realization from a spatial Gaussian process at point x. In the stochastic framework of model (3.1), with all else held equal, an increase in the value of u_t results in both an increase in the probability of observation $\xi(x,t)$ for all grid cells x, as well as, assuming $\alpha_+ \neq \alpha_-$, a change in the rate functions $\lambda_{mt}(x,t)$, which may translate to an increase or decrease in disease risk, depending on the signs of α_+ and α_-.

Temporal trends in the latent process are thus realized through specification of $U = (u_1, \ldots, u_T)$, where we assume there are T total time intervals of consideration in the analysis. Decreasing trends are brought about through specification of $u_t < u_{t+1}$, increasing through $u_t > u_{t+1}$, and alternating by $u_t < u_{t+1}$ for even t and $u_t > u_{t+1}$ for odd t. In this study time was discretized into 7 intervals of width 5 years. Thus, the U value associated with each temporal trend contains 7 elements.

For the increasing trend, $U = (-2.5, -2.0, -1.0, -0.5, 0.0, 0.5, 1.0)$, for the decreasing, $U = (1.0, 0.0, -0.5, -1.0, -2.0, -2.5, -3.0)$, and for the alternating,
$U = (-1.50, 0.00, -1.00, 0.5, -1.25, 1.00, -0.60)$, which are summarized in Table (3.1).

Trend	U
increasing	(-2.5, -2.0, -1.0, -0.5, 0.0, 0.5, 1.0)
decreasing	(1.0, 0.0, -0.5, -1.0, -2.0, -2.5, -3.0)
alternating	(-1.50, 0.00, -1.00, 0.5, -1.25, 1.00, -0.60)

Table 3.1: Parameters u_t used for each temporal trend simulated. For each level, the notational convention adopted here denotes $U = (u_1, \ldots, u_7)$, where u_t ($t \in \{1, \ldots, 7\}$) are the temporal parameter values at each time index.

To reflect realistic changes in environmental conditions over time, the fixed effect covariates $z_\lambda(x, t)$ vary by time interval t. For this simulation study these covariates were derived from the PRISM climatic dataset.

The PRISM dataset used here consists of a variety climatic measurements conducted at the 16 km^2 resolution averaged over a yearly time window. Specifically, these measurements consist of mean temperature, maximum temperature, minimum temperature, precipitation, minimum vapor pressure deficit, maximum vapor pressure deficit, and mean dew point temperature. Yearly averages of these quantities chosen from the years 1983, 1988, 1993, 1998, 2003, 2008, 2013, which were taken as the center years of the 7 time intervals (of width 5 years) which form the temporal extent of the simulations. For simplicity, the PRISM measurements from each center year were assumed to hold for all data simulated within the corresponding time interval. For each set of measurements from a given time interval, the PRISM values were not directly used as the covariates $z_\lambda(x, t)$, but rather, were first range standardized and then dimensionally reduced by principal component analysis. The range standardization was calculated as

$$r_{ikt} = (x_{ikt} - \min_{it})/(\max_{it} - \min_{it}) \text{ for } (i = 1, \ldots, 7), (t = 1, \ldots, 7)$$

where i indexes the measurement type, k indexes the raster cell in the study region for which the measurement was taken, t indexes time interval, $\min_i$ is the mini-

mum value of the ith measurement at year t and $\max_i$ is the maximum value of the ith measurement at year t. For each time interval, principal components of the 7 range standardized measurements were calculated using the rasterPCA function of the RStoolbox package, from the R programming language. The first 2 principal components (Figures 3.2, 3.3) corresponding to each time interval t where then used as covariates $z_\lambda(x,t)$. An intercept term was also included in $z_\lambda(x,t)$. The first principal component corresponds primarily to the temperature related variables, while the second is comprised mostly of moisture related variables. First (Figure 3.2) and second (Figure 3.3) principal components show subtle changes over time. Note that while the below figures show the principal components at a 16 km^2 resolution, for the three simulated datasets these principal components were used at a resolution of 2,458 km^2, out of consideration for speed of model fitting, corresponding to a discretization of the study region into 222 grid cells.

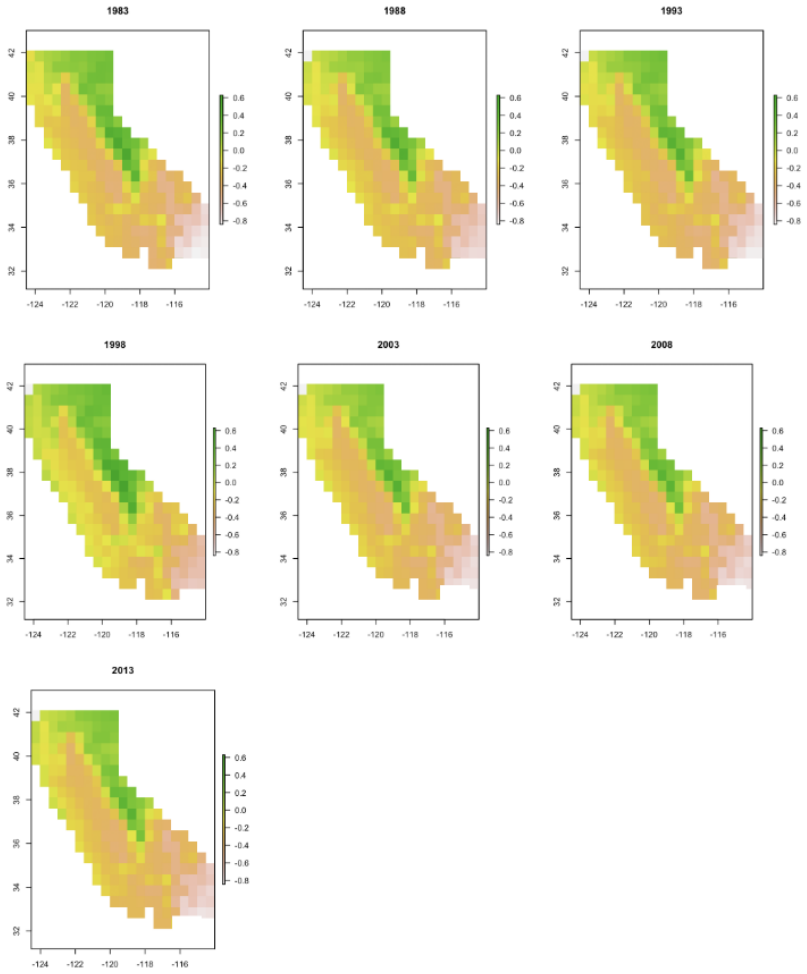

Figure 3.2: First principal components of the PRISM climatic dataset re-calculated over 7 different years between 1983 and 2013 at 2,458 km^2 resolution.

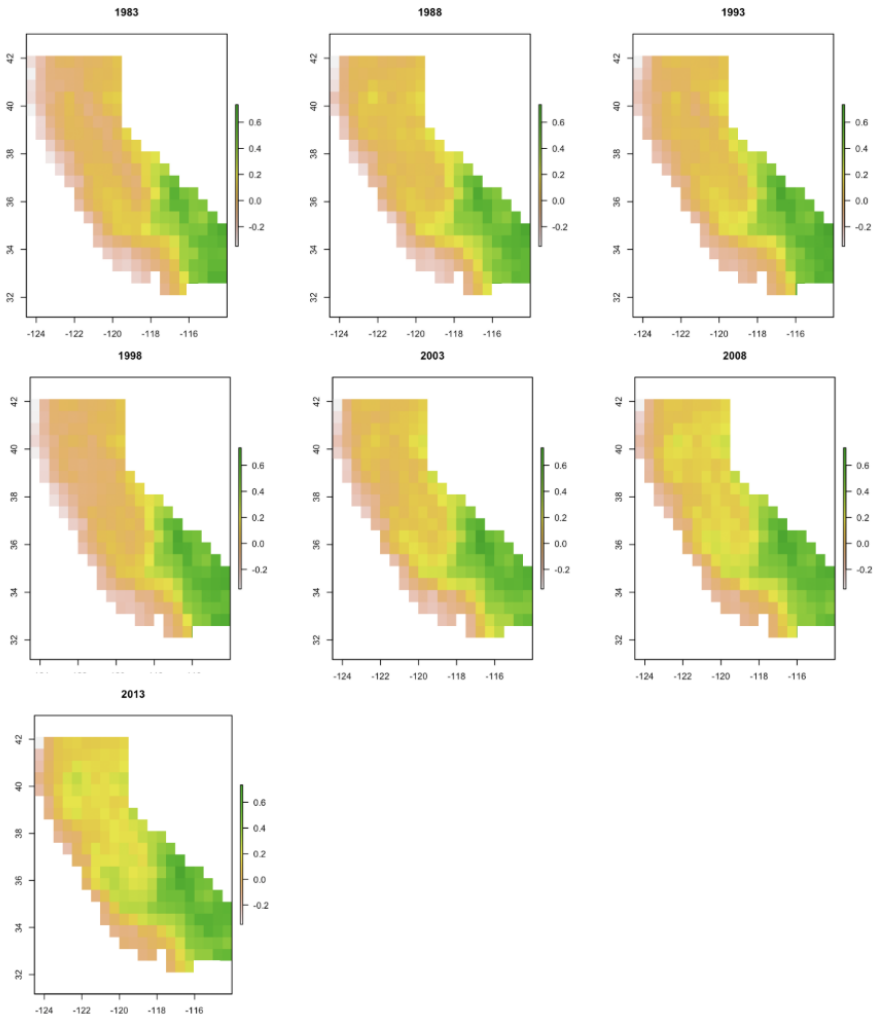

Figure 3.3: Second principal components of the PRISM climatic dataset re-calculated over 7 different years between 1983 and 2013 at 2,458 km^2 resolution.

The remaining parameters of model (3.1) were selected (Table 3.2) so as to yield mean disease prevalences between 0.081 and 0.449 (Table 3.3) for each temporal trend. For all datasets simulated, the spatial range θ and marginal variance ϕ of model (3.1) were set to $(\theta, \phi) = (7, 12)$

Trend	β_+	β_-	α_+	α_-
increasing	(0.5, 0.5, 0.5)	(4.5, 0.75, 0.75)	1	0.20
decreasing	(0.75, 0.50, 0.50)	(4.25, 0.75, 0.75)	1	0.20
alternating	(0.75, 0.50, 0.50)	(3.75, 0.75, 0.75)	0.75	0.20

Table 3.2: Simulation parameters used for each temporal trend. For all trends simulated, the spatial range θ and marginal variance ϕ of model (3.1) were set to $(\theta, \phi) = (7, 12)$.

Trend	1983	1988	1993	1998	2003	2008	2013
Increasting	0.081	0.104	0.169	0.203	0.252	0.300	0.351
Decreasing	0.449	0.331	0.279	0.233	0.164	0.137	0.110
Alternating	0.151	0.228	0.171	0.254	0.160	0.292	0.189

Table 3.3: Simulated mean disease prevalences per time interval. The second through final columns of this table refer to the central year around which the time interval was constructed.

Lastly, the mean numbers of observed grid cells per time interval, for each temporal trend, range from at least 52.56 to at most 144.04 (Table 3.4).

Trend	1983	1988	1993	1998	2003	2008	2013
Increasing	53.56	63.36	84.00	97.08	109.32	120.52	132.80
Decreasing	144.04	120.00	109.68	96.76	73.20	60.44	52.56
Alternating	75.92	109.36	85.88	120.48	79.80	131.20	96.84

Table 3.4: Mean numbers of simulated observation sites per time interval for each temporal trend. Each observation site corresponds to a single grid cell in the study region. For reference, the study area consists of 222 grid cells. The second through final columns of this table refer to the central year around which the time interval was constructed.

3.3.3 Results

For each of the 75 simulated datasets, the spatiotemporal model (3.1) was fit via a combination of Hamiltonian Monte Carlo (HMC) samplers, Metropolis-Hastings

random walk, and Gibbs sampling, as described in the methods section. The following prior distributions were specified:

$$
\begin{aligned}
\alpha_+ &\sim N(\hat{\alpha}_{+,initial}, 2) \\
\alpha_- &\sim N(\hat{\alpha}_{-,initial}, 2) \\
\beta_+ &\sim N(0, 100) \\
\beta_- &\sim N(0, 100) \\
\theta &\sim \text{Gamma}(shape, scale|\hat{\theta}_{initial}) \\
\phi &\sim \text{Inverse-Gamma}(shape, scale|\hat{\phi}_{initial}) \\
u_t &\sim N(0, 2) \text{ for } t = 1, \ldots, 7
\end{aligned}
$$

where $\hat{\alpha}_{+,initial}, \hat{\alpha}_{-,initial}, \hat{\theta}_{initial},$ and $\hat{\phi}_{initial}$ were obtained by fitting the modeling heuristic for parameter initialization described in the methods section of Project 1. In this case the heuristic model was fit to the temporally aggregated dataset. This additional step taken for parameter initialization did facilitate model convergence but was not strictly necessary for convergence to be reached. The shape and scale parameters pertaining to θ were chosen so that the prior mean of θ was equal to $\hat{\theta}_{initial}$, ie, the heuristically obtained initial estimate of θ, and the prior variance equal to 2. Similarly the shape and scale for ϕ were chosen so that the prior mean of ϕ equalled its initial estimate with a prior variance of 2. For each simulated dataset the MCMC routine of model (3.1) was run for a total of 8,000 iterations, with a burnin period of 3,000. The tuning periods for each parameter updated by HMC were set to 2,000 iterations, with target acceptance rates of 0.65.

The aggregated preferential sampling model was fit to the data in a nearly identical

fashion to that of the spatiotemporal model, save for the absence of updating the vector U of temporal parameters. For this model, parameters were also initialized by using the output of the simpler heuristic model, consisting of spatial logistic regression fit to the observational indicators κ_x in model (1.1). Posterior means of the vector of spatial random effects w along with θ and ϕ estimated by this spatial logistic model were used as initial values. Then, initial values for the remaining parameters were taken as the maximum likelihood estimates of the case and control data regressed upon the PRISM climatic principal components and the posterior mean estimate of w obtained from the previous step. Further details regarding this initialization process can be found in the methods chapter of Project 1. Lastly, the temporal Poisson model was fit by maximum likelihood estimation, specifically, using the glm function of the R programming language.

For all temporal trends, and for every time interval therein, the spatiotemporal model had lowest average root mean squared error in estimated log disease odds (Table 3.5). For the following summaries we adopt the notational convention for model reference which refers to estimates obtained from model (3.1) as "temporal", those from the temporal Poisson model as "GLM", and those from the aggregated preferential sampling model as "pooled". We see from Table (3.5) that the differences in RMSE can be quite pronounced. For instance, under the increasing temporal trend for the time interval with a central year of 1983, the spatiotemporal model had an average RMSE of 0.466 (standard deviation: 0.327), far below the mean values of 2.148 and 7.917 for the Pooled and GLM models, respectively. In general, the GLM model had the highest RMSE across all temporal trends and time intervals.

Trend	Model	Metric	1983	1988	1993	1998	2003	2008	2013
Inc	Temporal	Mean	0.466	0.440	0.442	0.436	0.433	0.429	0.428
Inc	Temporal	SD	0.327	0.279	0.296	0.281	0.282	0.271	0.273
Inc	Pooled	Mean	2.148	1.866	1.316	1.265	1.175	1.286	1.502
Inc	Pooled	SD	1.070	1.146	1.251	1.512	1.402	1.409	1.420
Inc	GLM	Mean	7.917	3.117	5.773	2.512	2.567	2.51	2.367
Inc	GLM	SD	22.458	1.599	15.693	0.915	0.899	0.917	0.812
Dec	Temporal	Mean	0.691	0.703	0.696	0.705	0.704	0.719	0.720
Dec	Temporal	SD	1.132	1.171	1.148	1.158	1.189	1.187	1.218
Dec	Pooled	Mean	1.543	0.960	0.915	1.080	1.600	2.013	2.464
Dec	Pooled	SD	0.727	0.643	0.607	0.581	0.518	0.471	0.441
Dec	GLM	Mean	2.455	2.588	2.600	2.615	3.885	2.936	3.717
Dec	GLM	SD	0.751	0.846	0.830	0.777	5.566	0.857	3.725
Alt	Temporal	Mean	0.416	0.409	0.403	0.400	0.411	0.393	0.405
Alt	Temporal	SD	0.703	0.684	0.675	0.666	0.702	0.646	0.678
Alt	Pooled	Mean	0.772	0.619	0.656	0.716	0.709	0.893	0.586
Alt	Pooled	SD	0.213	0.292	0.189	0.308	0.172	0.356	0.203
Alt	GLM	Mean	1.801	1.674	1.831	1.577	1.740	1.569	1.691
Alt	GLM	SD	0.598	0.501	0.724	0.443	0.483	0.443	0.496

Table 3.5: Means and standard deviations (SD) of RMSE in estimated log disease odds for each time interval for simulated increasing (Inc), decreasing (Dec) and alternating (Alt) temporal trends. The fourth through final columns of the table identify the center year of the time interval over which RMSE was calculated.

Inspection of the patterns in RMSE over time shows striking characteristics. For the increasing temporal trend, the "pooled" approach witness a general decrease in mean RMSE of estimated log disease odds over time (Figure 3.4). The spatiotemporal model maintains very low RMSE for all time intervals, with only a very minute decrease in RMSE as time increases. Similarly, the RMSE of the "pooled" approach tends to increase over time for the decreasing temporal trend, while that of the spatiotemporal model tends to remain low once more. Lastly, RMSE generally showed a slight alternating trend over time from all models when fit to the dataset with an alternating temporal trend in risk, with the exception of the spatiotemporal model, which had consistently low RMSE (Figure 3.4).

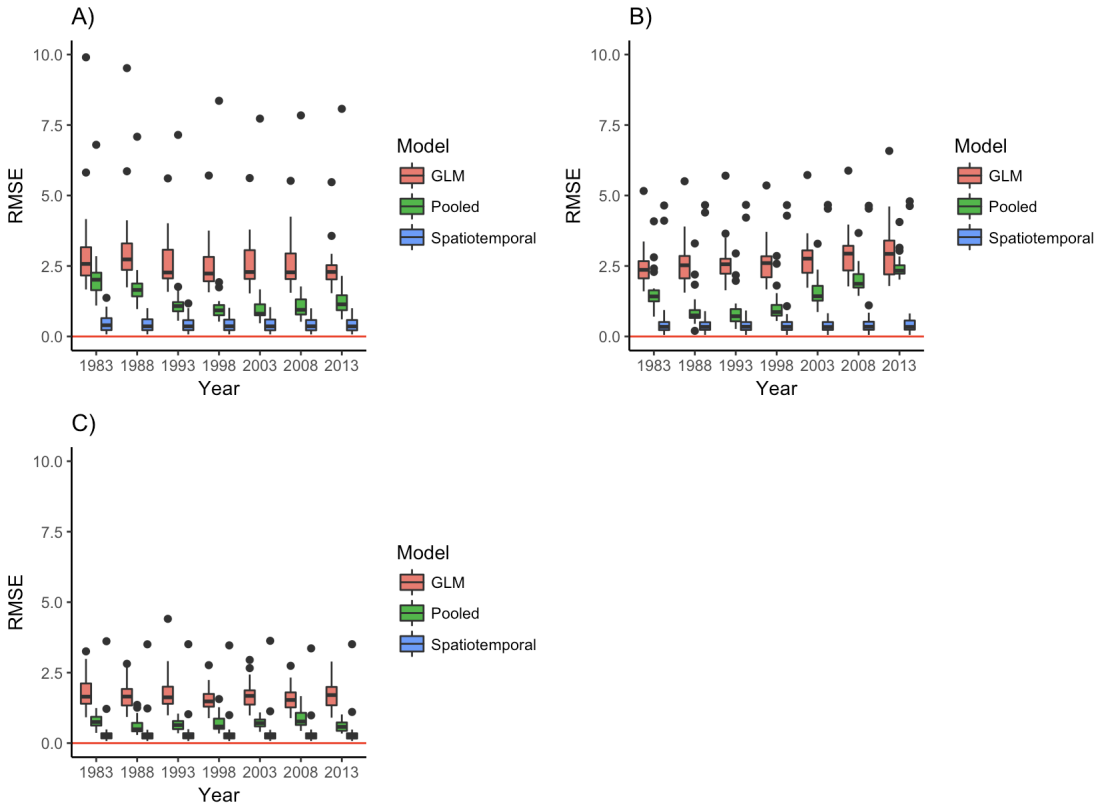

Figure 3.4: Boxplots of root mean squared errors in estimated log disease odds by year under A) increasing, B) decreasing, and C) alternating temporal trends for each evaluated model.

When inspecting RMSE in estimated log disease odds for the proposed spatiotemporal model in relation to the number of observed grid cells in the simulated datasets, we see that the few outliers in RMSE from this model are associated with datasets in which the number of observed cells tends to be low, particularly, under 75 (Figure 3.5).

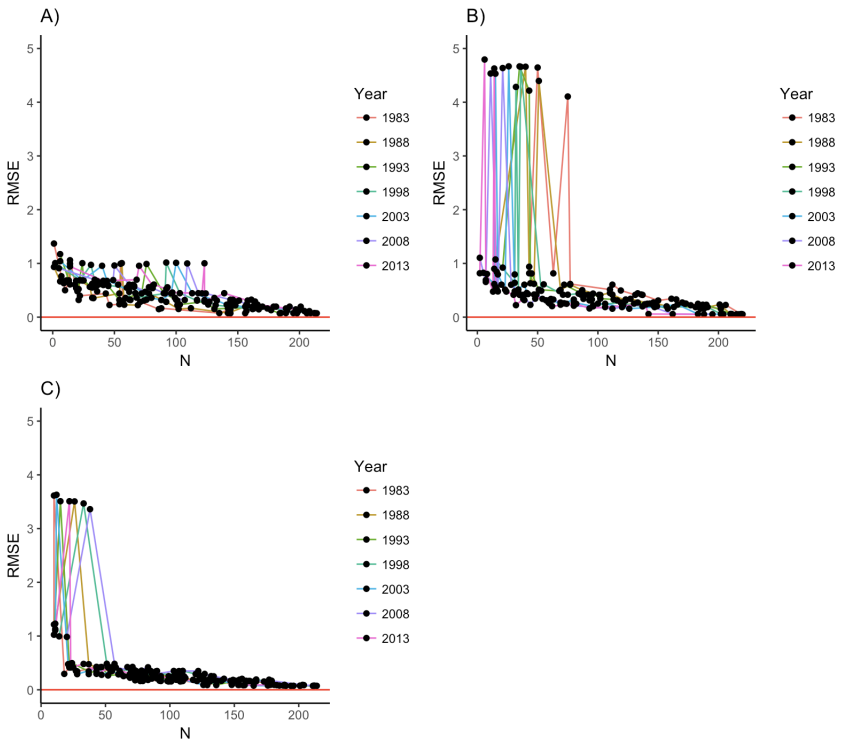

Figure 3.5: Root mean squared errors in estimated log disease odds of the proposed spatiotemporal model versus numbers of observed grid cells. Lines are color coded by the time interval over which RMSE was calculated. Thus, an increase in N for a given color code corresponds to an increase in the number of observed cells over the given window.

We now examine the convergence of model (3.1) for certain parameters of special interest, namely the spatial random effects w and the temporal parameters $u_1, \ldots, u_7$. Elevated average root mean squared errors in estimated spatial random effects were apparent, taking values of 2.446, 2.207, and 2.276 for increasing, decreasing, and alternating temporal trends, respectively (Table 3.6).

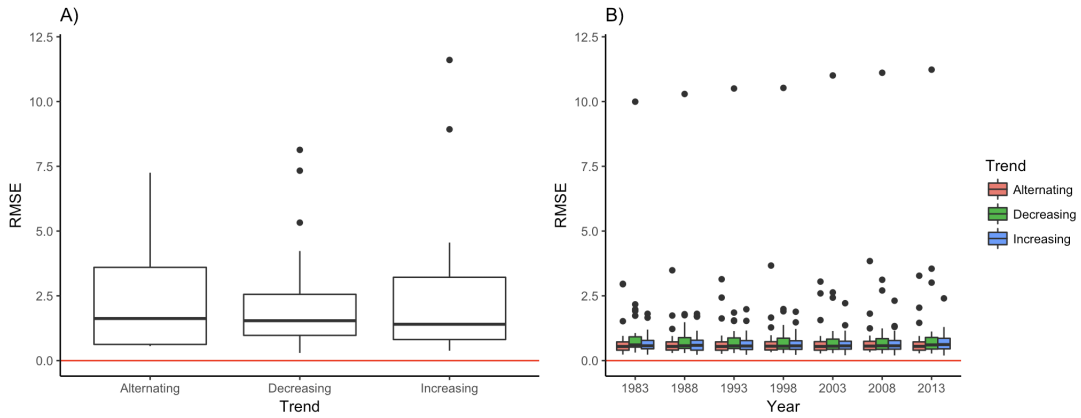

Figure 3.6: Summary of root mean squared errors in A) estimated spatial random effects $w(x)$ and B) estimated values of the spatiotemporal process $u_t + w(x)$ under differing temporal trends.

In contrast, average root mean squared errors in the sum $u_t + w(x)$ $(t = 1, \ldots, 7)$, remained low for all time intervals and temporal trends in the risk surface, reaching at most 1.249 for the decreasing trend in the time interval with central year 2013, and at minimum 0.656 for the alternating trend at the interval centered around 2013 (Table 3.7).

Trend	Mean	SD	Median	Q1	Q3
Increasing	2.446	2.668	1.402	0.813	3.216
Decreasing	2.207	2.046	1.539	0.974	2.557
Alternating	2.276	1.84	1.624	0.625	3.599

Table 3.6: Summary of root mean squared errors in estimated spatial random effects $w(x)$ from the proposed spatiotemporal model for differing temporal trends (SD: standard deviation, Q1: first quartile, Q3: third quartile).

Trend	Metric	1983	1988	1993	1998	2003	2008	2013
Increasing	Mean	0.687	0.700	0.670	0.671	0.684	0.708	0.718
Increasing	SD	0.398	0.449	0.410	0.396	0.423	0.442	0.443
Decreasing	Mean	1.194	1.123	1.112	1.116	1.162	1.208	1.249
Decreasing	SD	1.910	1.956	1.999	2.011	2.131	2.173	2.218
Alternating	Mean	0.764	0.731	0.756	0.749	0.757	0.765	0.741
Alternating	SD	0.714	0.659	0.683	0.685	0.683	0.721	0.656

Table 3.7: Summary of mean and standard deviation (SD) in root mean squared errors of the estimated spatiotemporal process, $w(x) + u_t$, obtained from model (3.1) for differing temporal trends.

Lastly, we consider biases in the estimates of the spatial and preferential sampling parameters in model (3.1). Low average biases were observed in the preferential sampling parameters of cases and controls, α_+ and α_-, remaining below a magnitude of 0.15 for all temporal trends. Bias in the spatial range θ also was limited, reaching at most an average of 1.056 (standard deviation: 4.626) for the decreasing temporal trend, while that of the marginal variance ϕ reached a maximum of only 0.393 on average (Table 3.8).

Trend	Metric	α_+	α_-	θ	ϕ
Increasing	Mean	-0.033	-0.005	-0.019	0.079
Increasing	SD	0.137	0.027	1.348	0.079
Decreasing	Mean	-0.128	-0.028	1.056	0.393
Decreasing	SD	1.246	0.257	4.626	0.393
Alternating	Mean	-0.102	-0.039	0.339	-0.239
Alternating	SD	0.718	0.256	1.370	0.839

Table 3.8: Summary of means and standard deviations in biases of estimates for additional parameters in model (3.1). α_+ and α_- are preferential sampling related parameters while θ and ϕ are the range and marginal variance of the latent spatial process, respectively.

3.3.4 Discussion

This simulation study examined the comparative performance of 3 different modeling approaches when fit to data simulated under a range of temporal trends in the latent

risk surface. For all temporal trends considered here, that is, increasing, decreasing, and alternating, the proposed spatiotemporal model substantially outperformed the reference methods in terms of root mean squared error in estimated log disease odds. Interestingly, the Poisson model performed the worst for all temporal trends, despite the fact that it should have accounted for the temporal changes within the data due to the fact that it entailed fitting separate Poisson models to the simulated data within each time interval. This result serves as a warning that accounting for temporal trends while ignoring the preferential sampling mechanism behind the data can still lead to deteriorated quality in estimated disease odds.

The change in RMSE of log disease odds over time shows a striking trend worthy of explanation. When the models were fit to data simulated under an increasing temporal trend in the latent risk surface, the RMSE tended to decrease on average as time increased. Under a decreasing temporal trend, the RMSE tended to increase with time, while under an alternating trend, the RMSE alternated with time. This observed phenomenon can ultimately be explained by changes in the volume of data simulated over time. We recall that under our scenario of preferential sampling, the distribution of observation sites is stochastically related to the risk of the underlying disease. As disease risk increases over space or over time, so too should the number of observation sites increase. We see this tendency manifested in Table (3.4), which, for the increasing trend in disease risk, shows an increase in the average number of observation sites from the first to last interval. This increase in observation sites coincides with an increase in the total number of collected specimen on average (cases + controls). Consequently, due to this greater volume of data in later years, both in terms of the numbers observation sites as well as cases and controls, the later time intervals exert a greater influence on the estimation of the spatial random field w in model (1.1). Since the true w changes over time (as $u_t + w(x)$ in model (3.1)), the estimated w, assumed constant in the aggregated model, becomes biased towards

its true value in later years. In this way also, estimates of log disease odds become biased toward later time intervals, in which they have lower RMSEs, and become biased away from earlier intervals, where they yield higher RMSEs. The interplay between underlying disease risk, the number of observation sites, and the direction of bias works in the opposite fashion when the underlying data possess a decreasing temporal trend. Similarly, under the alternating trend, estimates become biased toward years which witness higher disease risk.

We now turn to the estimability of specific parameters in model (3.1), namely the vector of spatial random effects w and the temporal parameters u_t. Despite the failure of the model to converge to the random effect values w, convergence was achieved for the sum $u_t + w(x)$. From this we conclude that w and u_t are not identifiable from the data, but the sum $u_t + w(x)$ is. Given that it is only the spatiotemporal process $u_t + w(x)$ which influences estimation of disease risk, and not the values of u_t or $w(x)$ individually, we view this lack of identifiability as not a serious impediment to the use of the spatiotemporal model in practice. Regarding the remaining parameters of model (3.1), we observe that low bias estimates are obtained from our fitting procedure for all temporal trends considered here. Thus, we conclude that the proposed model substantially outperforms reference methods for reasonable sample sizes and temporal trends in risk, and adequately models these temporal trends in practice, even if the specific parameters u_t and w are not estimable in isolation.

3.4 Analysis

3.4.1 Introduction

This analysis applies the proposed spatiotemporal preferential sampling method to a disease surveillance dataset obtained from the California Department of Public Health (CDPH), targeting infection with *Yersinia pestis*, or plague, in the Sciurid population of California. Sciurids, or the rodent family of squirrels, are susceptible to plague at a prevalence of roughly 6% in California. This low but persistent prevalence, coupled with the fact that many human cases of plague are linked to epizootics in the animal population, make Sciurids an ever-relevant focus of disease surveillance.

We recall that Project 1 of this dissertation provided an analysis of this dataset with a model to correct for preferential sampling, or the stochastic dependency between disease risk and the distribution of sampling sites. However, the approach therein was limited in the sense that it aggregated observations throughout time over a window between 1983 and 2015. In many real world disease surveillance applications prominent temporal changes in the observed disease risk may arise due to a multitude of factors, such as fluctuations in climate or other abiotic factors, temporal trends in the underlying ecology of the host organisms, or budgetary forces impacting the distribution and extent of surveillance efforts. In short, aggregating data over time may obscure key dynamics of both the underlying disease as well as the observation process. To offer a more sophisticated treatment of temporal trends, we now apply a newly developed spatiotemporal extension of the original preferential sampling model. Our core objective is thus to determine whether modeling the sampling process over time can capture meaningful temporal trends in predicted risk that were overlooked by the time-aggregated approach.

The CDPH surveillance system targets plague in the rodent family of squirrels, known as *Sciuridae* or *Sciurids*, which in this dataset comprise 21 different species, namely the: Antelope Ground Squirrel, Antelope Ground Squirrel (WhiteTail), Belding's Ground Squirrel, California Ground Squirrel, Chipmunk, Least Chipmunk, Long-eared Chipmunk, Lodgepole Chipmunk, Merriam's Chipmunk, Panamint Chipmunk, Shadow Chipmunk, Siskiyou Chipmunk, Sonoma Chipmunk, Uinta Chipmunk, Yellow-pine Chipmunk, Golden-mantled Ground Squirrel, Ground Squirrel, Yellow-bellied Marmot, Pine Squirrel, and Squirrel. The surveillance system collects data by conducting a series of sampling events at locations throughout California in which Sciurids are trapped and subsequently tested for *Yersinia pestis* via F1 antigen tests. The data contain samples collected between 1983 and 2015. The surveillance system tends to assign sampling locations to high risk or high impact areas, where risk is assessed to be high in what are viewed as plague endemic regions, as determined by historic cases of plague in humans or recovered Sciurid specimen, and high impact areas are regions where human-Sciurid interactions are particularly likely, such as in national parks. In this sense the data are preferentially sampled, due to the stochastic relationship between the distribution of observation sites and the risk of the disease being surveilled.

In this analysis we fit the proposed model (3.1) against a benchmark, taken to be the preferential sampling model of Project 1, which aggregates the disease surveillance data over time. In this model we replace the time specific intervals κ_{xt} of model (3.1) with time-aggregated indicators, κ_x, which reflect whether the grid cell with center at point x contains any sampling events between 1983 and 2015. We also replace time specific counts Y_{xmt} with counts Y_{xm}, which have been aggregated over time. Lastly, the latent process shared between locational and disease related components of the model is merely spatial, rather than spatiotemporal.

As explained in the methods section, log disease odds for each time interval are given by the differences in the rate functions of cases and controls, $\log(\lambda_{+,t}(x))-\log(\lambda_{-,t}(x))$. Disease risk over time, the ultimate quantity of concern for our analysis, relates to log odds as $r_t(x)/[1-r_t(x)] = \lambda_{+,t}(x)/\lambda_{-,t}(x)$, where we denote the risk for plague in grid cell x over time interval t as $r_t(x)$. Similarly to the spatiotemporal model, the time-aggregated log disease odds are calculated as the differences of case and control rate functions, $\log(\lambda_{+}(x))-\log(\lambda_{-}(x))$ from model (1.1). Consequently, the risk calculated from model (1.1) relates to log disease odds via $r(x)/[1-r(x)] = \lambda_{+}(x)/\lambda_{-}(x)$, where $r(x)$ is the risk of plague in the grid cell centered at point x averaged over all years of observation between 1983 and 2015. Fitting both models (3.1) and (1.1) is thus intended to offer a comparison to illustrate the effect on predicted risk, $r_t(x)$ and $r(x)$ in the above notation, brought about by accounting for temporal changes in the surveillance data.

3.4.2 Data

The disease surveillance dataset considered in this analysis consists of observations conducted in the state of California between 1983 and 2015, recording the tested presence or absence of plague in Sciurids. Observations in this dataset have been collected continuously on a yearly basis, with no year between 1983 and 2015 failing to have recorded data. In this study we utilized observations for all 21 species of Sciurids combined, producing a series of maps indicating the overall risk of plague in Sciurids as a whole, which we note does not necessarily equate to the risk of plague in individual species within the Sciuridae family. To fit models (3.1) and (1.1), the study region was discretized into 584 nonoverlapping grid cells of area 836 km^2. The timespan of the study was discretized into 7 windows of width 5 years, where the years 1983, 1988, 1993, 1998, 2003, 2008, and 2013 mark the center points of each of

the 7 time windows.

The temporally varying fixed effects of model (3.1), $z_\lambda(x,t)$, were derived from the publicly accessible PRISM climatic dataset, maintained by the Oregon State University. The PRISM dataset used here consists of a variety climatic measurements averaged over yearly time windows. Specifically, these measurements consist of mean temperature, maximum temperature, minimum temperature, precipitation, minimum vapor pressure deficit, maximum vapor pressure deficit, and mean dew point temperature. Yearly averages of these quantities for the years 1983, 1988, 1993, 1998, 2003, 2008, and 2013 were obtained. For simplicity, the PRISM measurements from each of these years were assumed to hold for all surveillance data collected within the corresponding time interval. For instance, disease surveillance observations conducted in the year 1989 were associated with the PRISM yearly averages for 1988, which marks the center point of the 5 year time window which forms the basic unit of temporal analysis. For each set of climatic measurements from a given time interval, the PRISM values were not directly used as the covariates $z_\lambda(x,t)$, but rather, were first range standardized and then dimensionally reduced by principal component analysis. The range standardization was calculated as

$$r_{ikt} = (x_{ikt} - \min_{it})/(\max_{it} - \min_{it}) \text{ for } (i = 1, \ldots, 7), (t = 1, \ldots, 7)$$

where i indexes the measurement type, k indexes the raster cell in the study region for which the measurement was taken, t indexes time interval, $\min_i$ is the minimum value of the ith measurement at year t and $\max_i$ is the maximum value of the ith measurement at year t. For each time interval, principal components of the 7 range standardized measurements were calculated using the rasterPCA function of the RStoolbox package, from the R programming language. The first 2 principal components corre-

sponding to time interval t where then used as covariates $z_\lambda(x,t)$. An intercept term was also included in $z_\lambda(x,t)$. The first principal component corresponds primarily to the temperature related variables, while the second is comprised mostly of moisture related variables. First (Figure 3.7) and second (Figure 3.8) principal components show subtle changes over time. Note that while the principal components are portrayed in (Figure 3.7) and (Figure 3.8) at the $16km^2$ resolution, the actual MCMC routine in which model (3.1) was fit used the principal components at an 836 km^2 resolution, that is, the resolution to which the study region was discretized. However, the finer resolution risk map obtained from model (3.1) was spatially downscaled to resolution of $16km^2$, yielding the high resolution risk map which is ultimately of greater concern to our analysis.

In contrast, the time-aggregated reference model did not incorporate the above temporally varying PRISM principal components. Instead, covariates $z_\lambda(x)$ from model (1.1) were taken to be the first two principal components of the PRISM 30 year average normals, rather than yearly values, for mean temperature, maximum temperature, minimum temperature, precipitation, minimum vapor pressure deficit, maximum vapor pressure deficit, and mean dew point temperature, or the same measurement types as were used in the construction of the covariates for the temporal model.

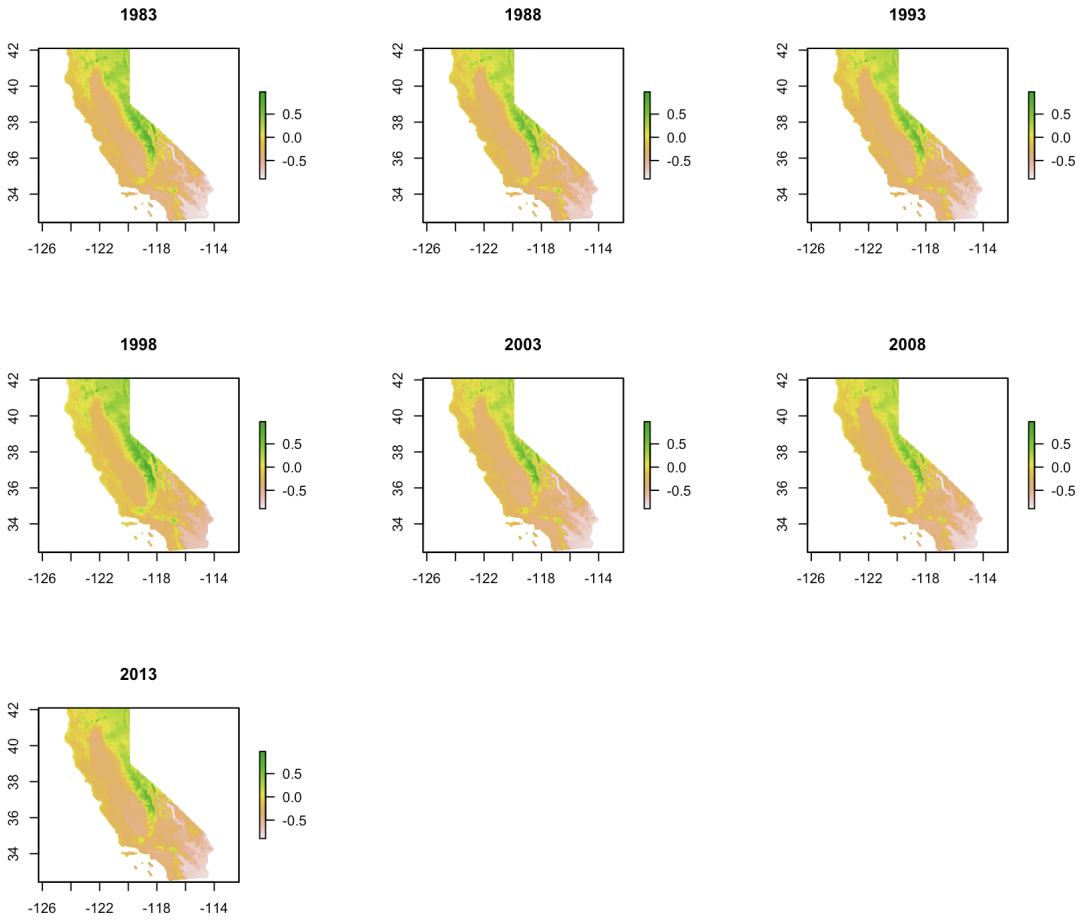

Figure 3.7: First principal components of the PRISM climatic dataset re-calculated over 7 different years between 1983 and 2013 at 16 km^2 resolution.

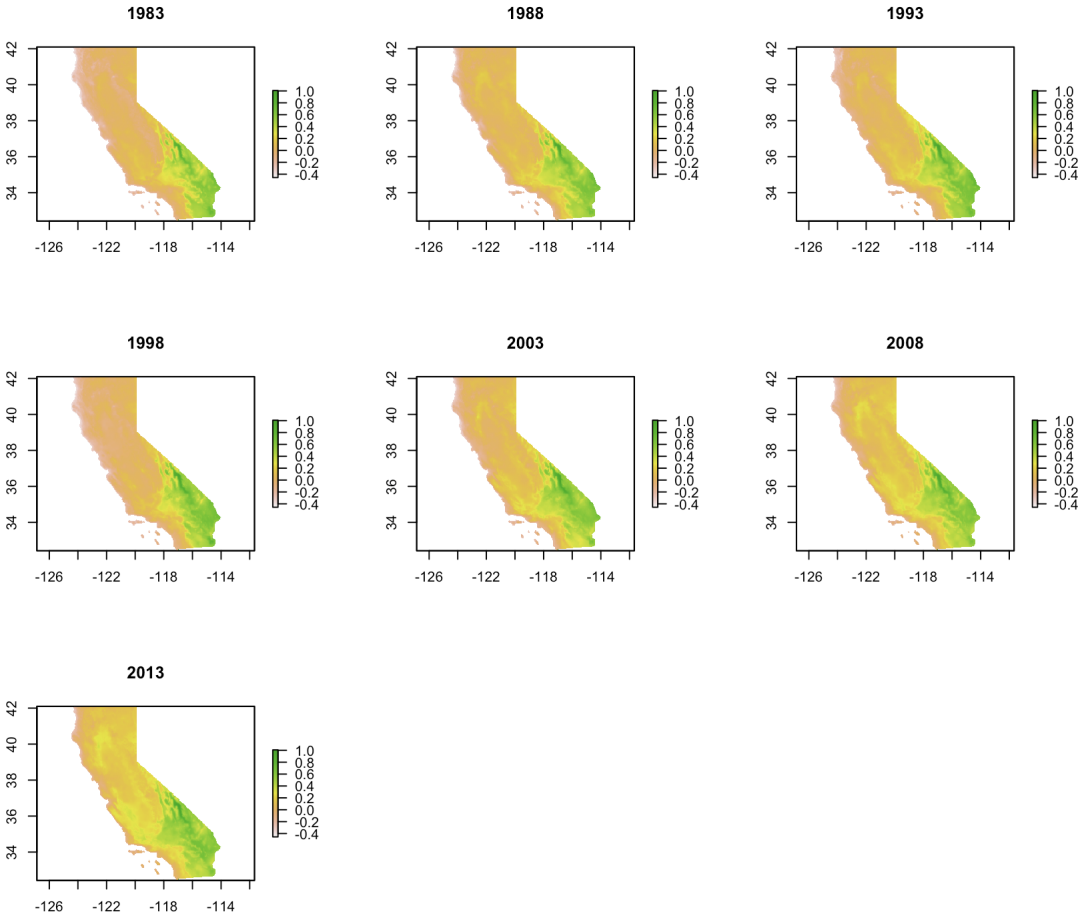

Figure 3.8: Second principal components of the PRISM climatic dataset re-calculated over 7 different years between 1983 and 2013 at 16 km^2 resolution.

The raw prevalence of plague shows notable trends over the 7 time windows of the study, ranging from a high of 0.11 in the time window centered at 1993 to a low of 0.02 in the window centered at 2003 (Table 3.9). In addition, the number of observed grid cells varies strongly over time. We recall that model (3.1) encodes the distribution of observation sites in terms of indicator variables $\kappa_{xt} \in \{0, 1\}$, which represent whether the grid cell with center point at location x contains at least 1 plague sampling event during time window t. Consequently, a decline in the number of observed grid cells would reflect a decrease in the observed spatial extent of the study region over time. The sum of κ_{xt}, that is, the total number of observed grid cells, declines steadily from a maximum of 114 in 1983 to a minimum of 64 in 2013 (Table 3.9).

Time Interval	Observed Cells	Total Specimen	Prevalence
1983	114	2513	0.08
1988	108	3931	0.08
1993	113	4119	0.11
1998	87	3311	0.05
2003	81	2947	0.02
2008	83	2499	0.03
2013	64	2447	0.05

Table 3.9: Summary of the total number of observed grid cells, total recorded specimen (cases + controls), and disease prevalences per time interval for Sciurids.

3.4.3 Results

The proposed spatiotemporal model was fit to the data with a total of 11,000 MCMC samples, converging after a burnin of 900 samples, while the reference, time-aggregated model (1.1) was fit with 10,000 MCMC samples and a burnin of 3,000. For both models, Hamiltonian Monte Carlo step sizes were self-tuned by the scheme of dual averaging to achieve target acceptance rates of 0.65, as calculated over tuning periods of 2,000 samples.

The temporally referenced risk maps obtained from model (3.1) are presented in

(Figure 3.9). In these risk maps the raster values represent the probability a rodent sampled at a particular location within a particular time interval will be plague positive. At first glance notable temporal trends appear in the estimated risk surface. Over time, the band of elevated risk stretching from roughly the 40th to 36 parallels, corresponding to the Sierra Nevada mountain range, steadily decreases in size, with an exception for the time interval centered in 1998. In addition, the small bands of elevated risk in the southwestern portion of the map, near Los Angeles and San Bernardino counties, appear to shrink as well. However, the overall macroscopic structure of the risk maps remains fairly consistent. The highest areas of elevated risk consistently arise in the Sierra Nevada mountains, while the southeastern region of the map maintains the lowest predicted risk.

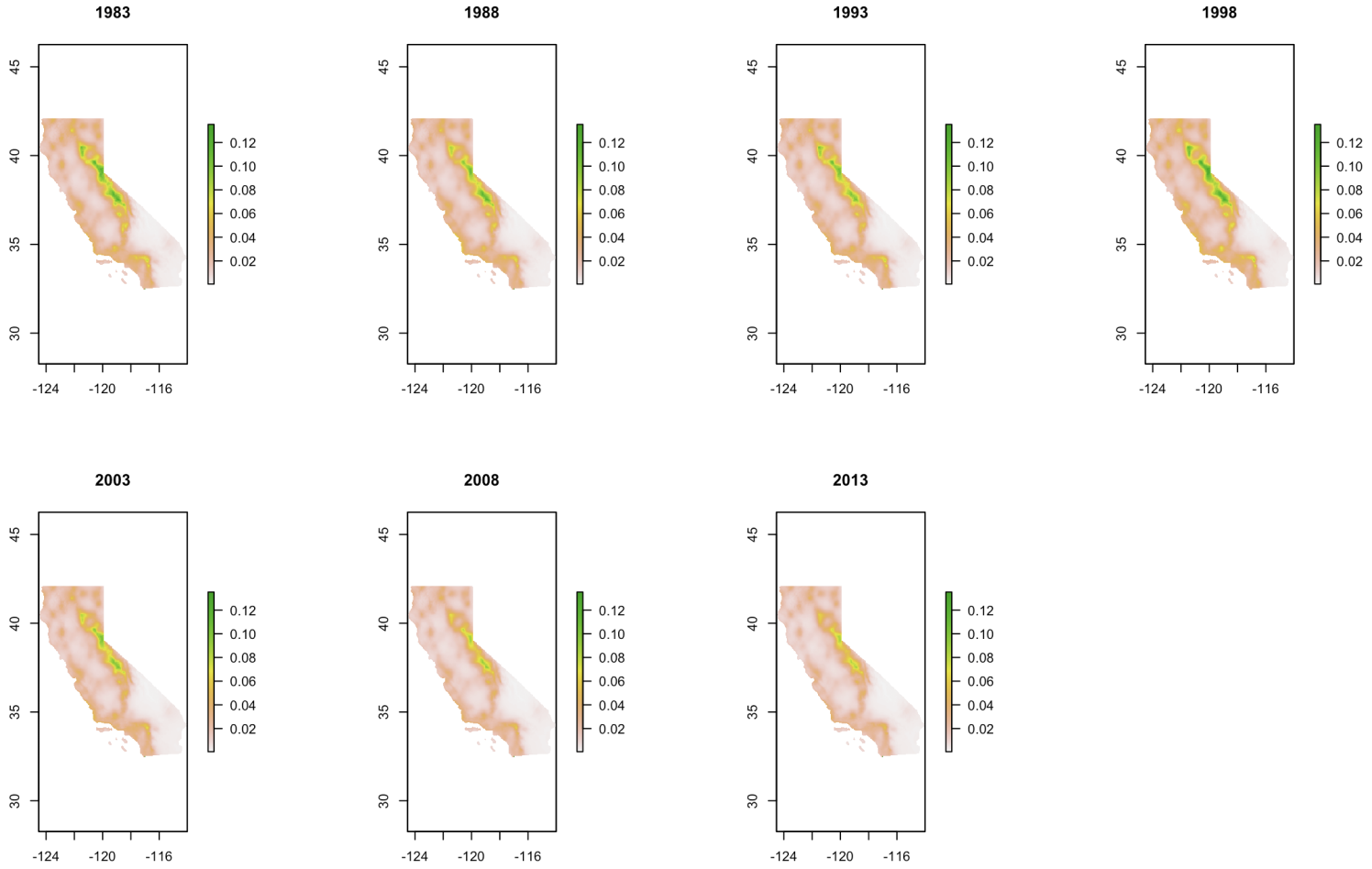

Figure 3.9: Estimated risk of plague in Sciurids for each time interval of the study, at a resolution of $16km^2$, as calculated by model (3.1).

The high resolution risk map estimated by the reference, time-aggregated model (1.1) is presented for comparison in Figure (3.10). In this risk map, as before, the raster values represent the probability a rodent sampled at a particular location will be plague positive, but now over the entire window of time between 1983 and 2015. This risk map ranges in value from 0.008 to 0.115. As with the spatiotemporal model, peak areas of risk fall along the Sierra Nevada mountain range, stretching diagonally from roughly the 40th to 35th latitude towards the eastern border of the state, as well as within thin pockets of elevated risk in the northeastern portion of the state, in addition to circular regions of greater risk towards the southwestern part of the map. The southern and central coastlines also show mild elevation in risk relative to some of the lower risk regions of the map, such as the San Joaquin Valley, to the west of the Sierra Nevada mountains, and the Imperial Valley region, in the southeastern corner.

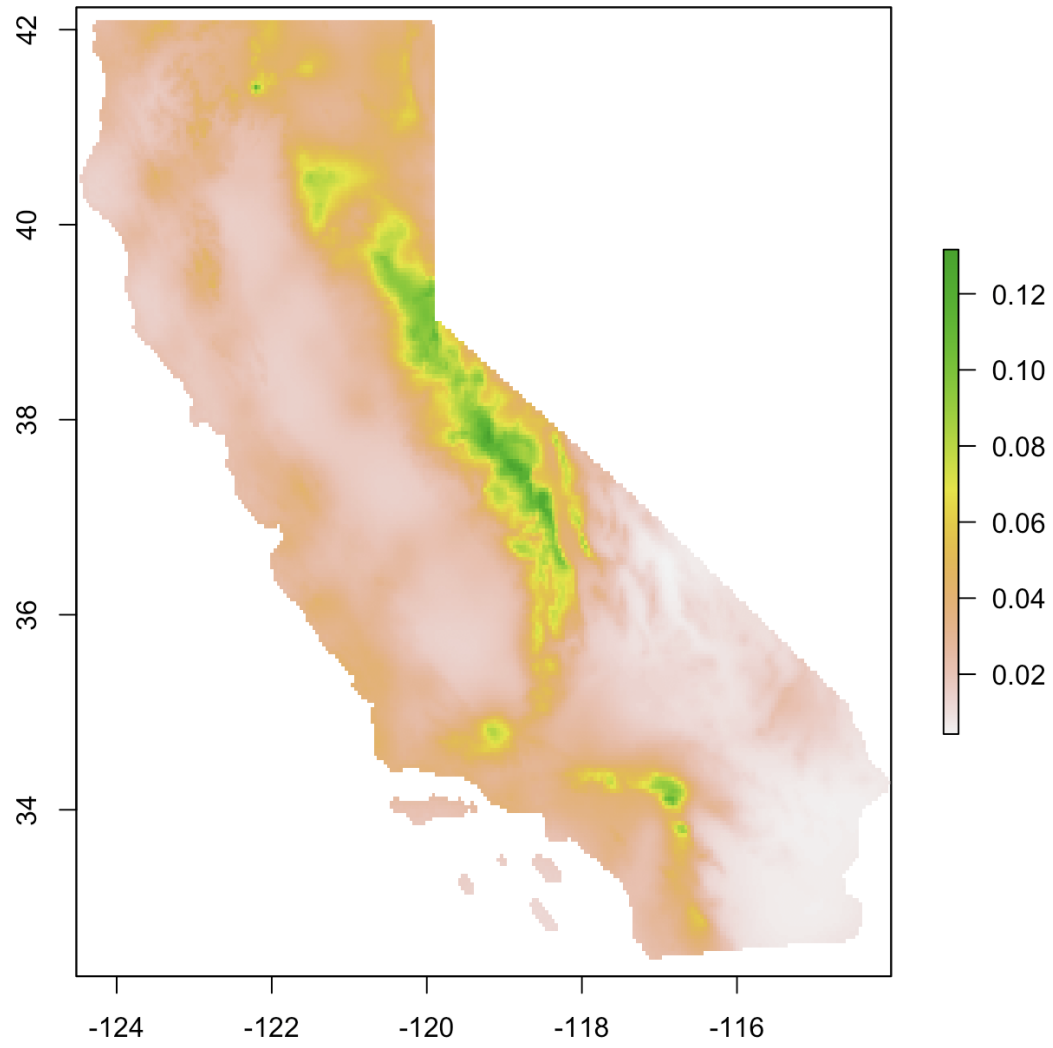

Figure 3.10: Risk of plague in Sciurids over all years between 1983 and 2015, at a resolution of $16km^2$, as calculated by the reference model (1.1).

The above maps offered a rough picture of the temporal trend in risk, showing a general decrease from 1983 to 2015, with the exception of the time intervals centered at 1998. For a more direct examination of the spatiotemporal trend we consider changes in $\hat{u}_t + \bar{\hat{w}}$ over time, where $\hat{u}_t$ are our posterior mean estimates of the time specific parameters of model (3.1) and $\bar{\hat{w}}$ is the average of our estimates for the spatial random effects

$$\bar{\hat{w}} = \frac{1}{584}\Sigma_{i=1}^{584}\hat{w}(x_i)$$

In other words, $\hat{u}_t + \bar{\hat{w}}$ is the estimated mean of the estimated latent spatiotemporal process during time interval t. We see that this quantity shows a general increase from the 1st to 4th time intervals (1983 to 1998), followed by a strong decrease until the 6th interval (2008), after which a slight increase appears (Figure 3.11A). To ascertain the driving factors behind this trend we also consider changes in the disease prevalence and number of observation sites over time (Figure 3.11B, C). In the construction of model (3.1) we recall that $u_t + w(x)$ relates to the distribution of observation sites as well as the observed abundances of cases and controls, and consequently, disease risk. Hence, variation in either prevalence or the number of observation sites may be expected to relate to $u_t + w(x)$. Both prevalence and the number of observation sites show a noted decline starting in the time interval indexed by 1993, which anticipates the decrease in $\hat{u}_t + \bar{\hat{w}}$ in the next time interval. The slight uptick in $\hat{u}_t + \bar{\hat{w}}$ arising in the last time interval coincides with a slight increase in disease prevalence, but not the number of observation sites. Thus, a fairly consistent relationship seems apparent between $\hat{u}_t + \bar{\hat{w}}$ and disease prevalence, as well as, to a lesser extent, the number of observation sites, with the exception of the one-interval lag between the decrease in the former beginning after 1998 and the decreases in the latter occurring after 1993.

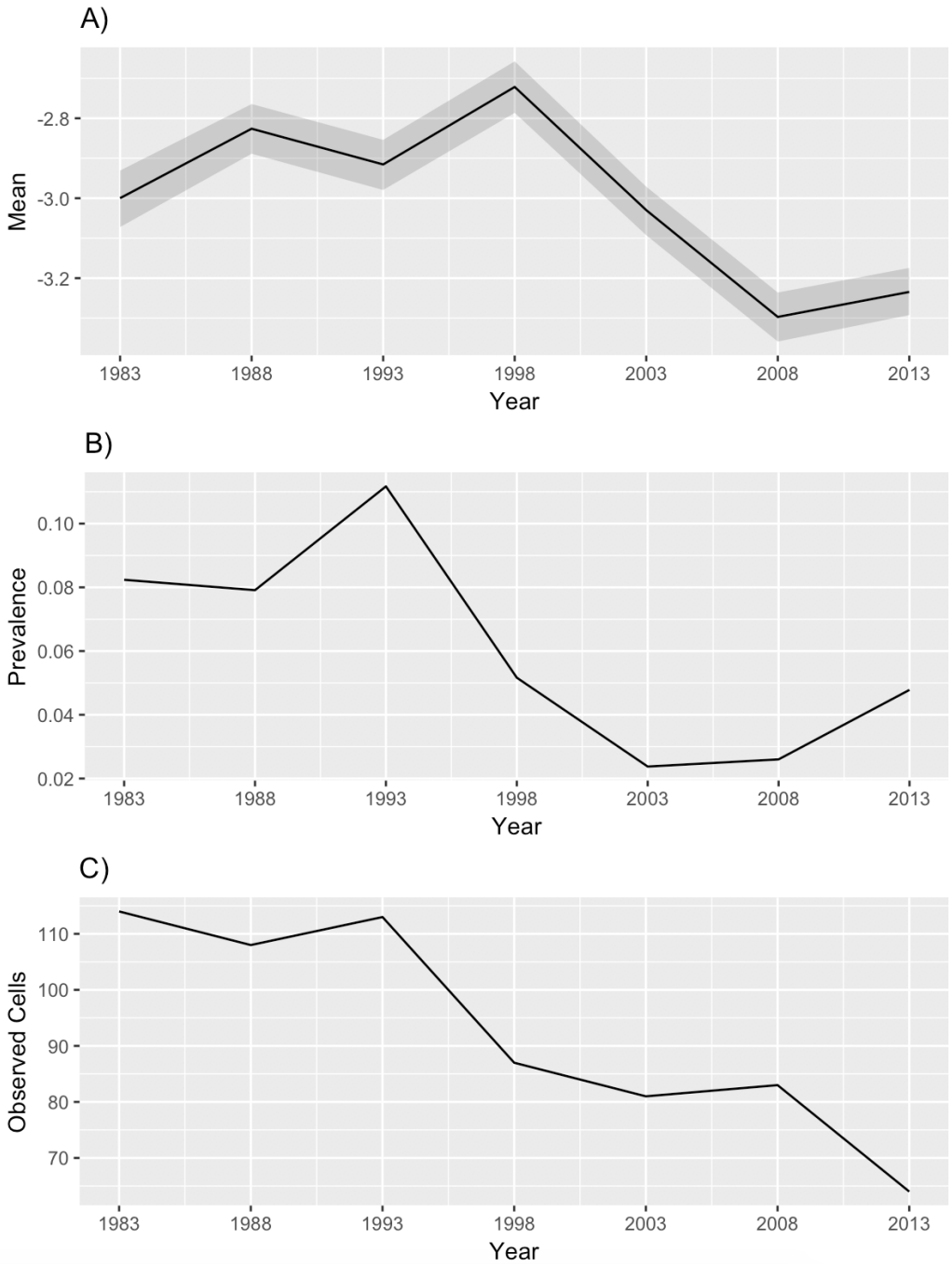

Figure 3.11: Comparison of A) estimated mean of the spatiotemporal process $u_t + w(x)$, with 25th and 75th posterior quantiles shaded B) disease prevalence over each time interval between 1983 and 2015 and C) the number of observed grid cells over each interval.

Returning back to the comparison of the spatiotemporal and time-aggregated models, we now examine per-cell differences in estimated risk (Figure 3.12). The temporal model tends to have lower estimated risk than the aggregated model, differing on average, -0.007, -0.008, -0.009, -0.004, -0.01, -0.013 and -0.014 for the 7 time intervals centered between 1983 and 2013. We observe that this average difference tends to increase in magnitude over time.

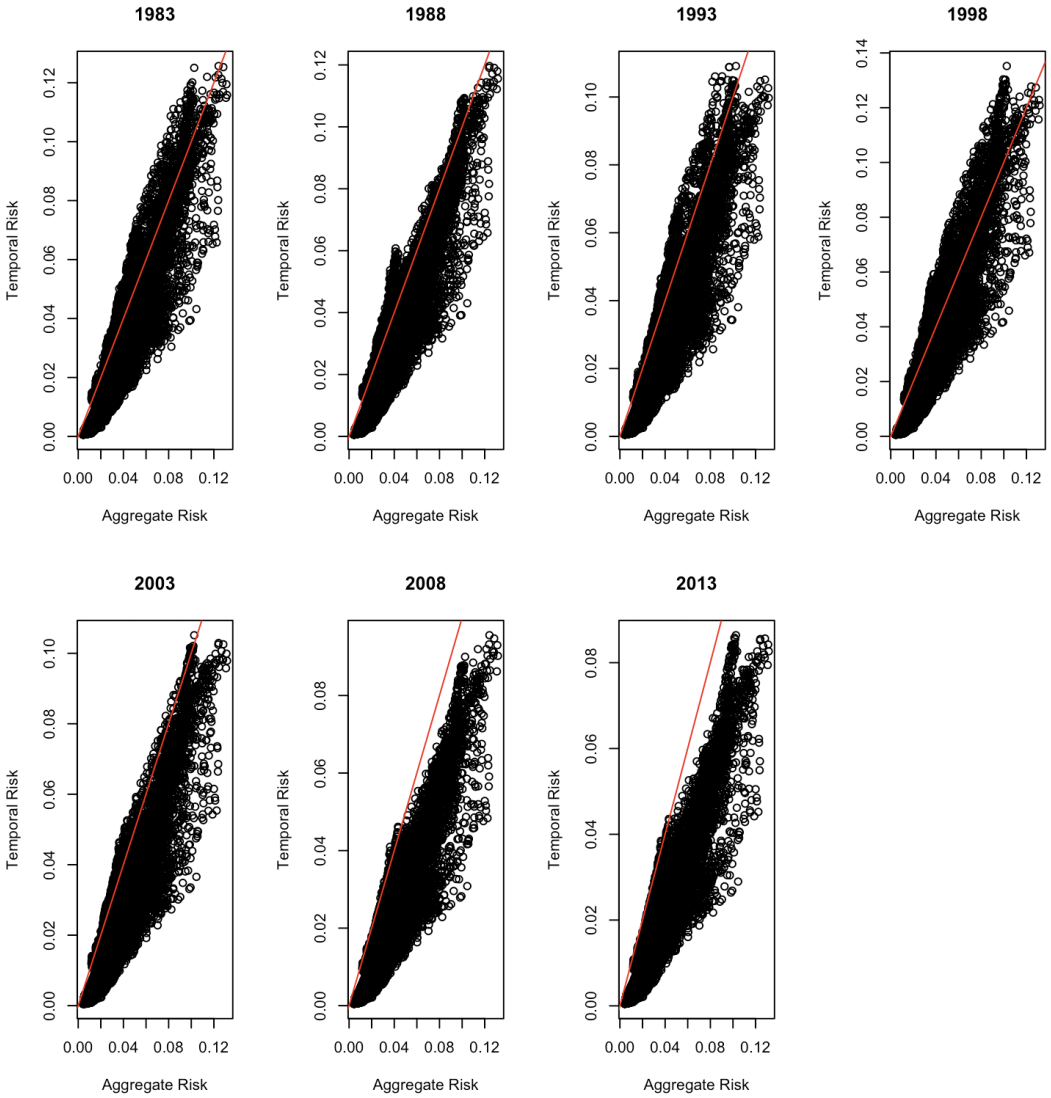

Figure 3.12: Scatterplots of risk calculated from the spatiotemporal model versus the time-aggregated model for each time interval of the study. The red lines are of slope 1 and intercept 0.

However, it is of greater interest to determine the degree to which the spatiotemporal and temporal models yield statistically significantly different estimates of risk. To that end, for each model we calculate the posterior probability of risk exceeding 0.05 over all pixels in the study region, as described in the methods section. Observing only pixels for which this probability exceeds 0.95, we find that the spatiotemporal model shows a decrease in areas of significant risk over time (Figure 3.13), similar to the trend noted when observing the values of estimated risk (Figure 3.9). This spatial coverage of significant risk is less than that of the time-aggregated model (Figure 3.14). In particular, the number of all pixels deemed significant by the spatiotemporal model is always less than that of the time-aggregated model. Fractions of the number of pixels deemed significant by model (3.1) divided by that deemed significant by model (1.1) take values of 0.702, 0.554, 0.556, 0.802, 0.459, 0.262 and 0.199 for the time intervals centered between 1983 and 2013 (Table 3.10).

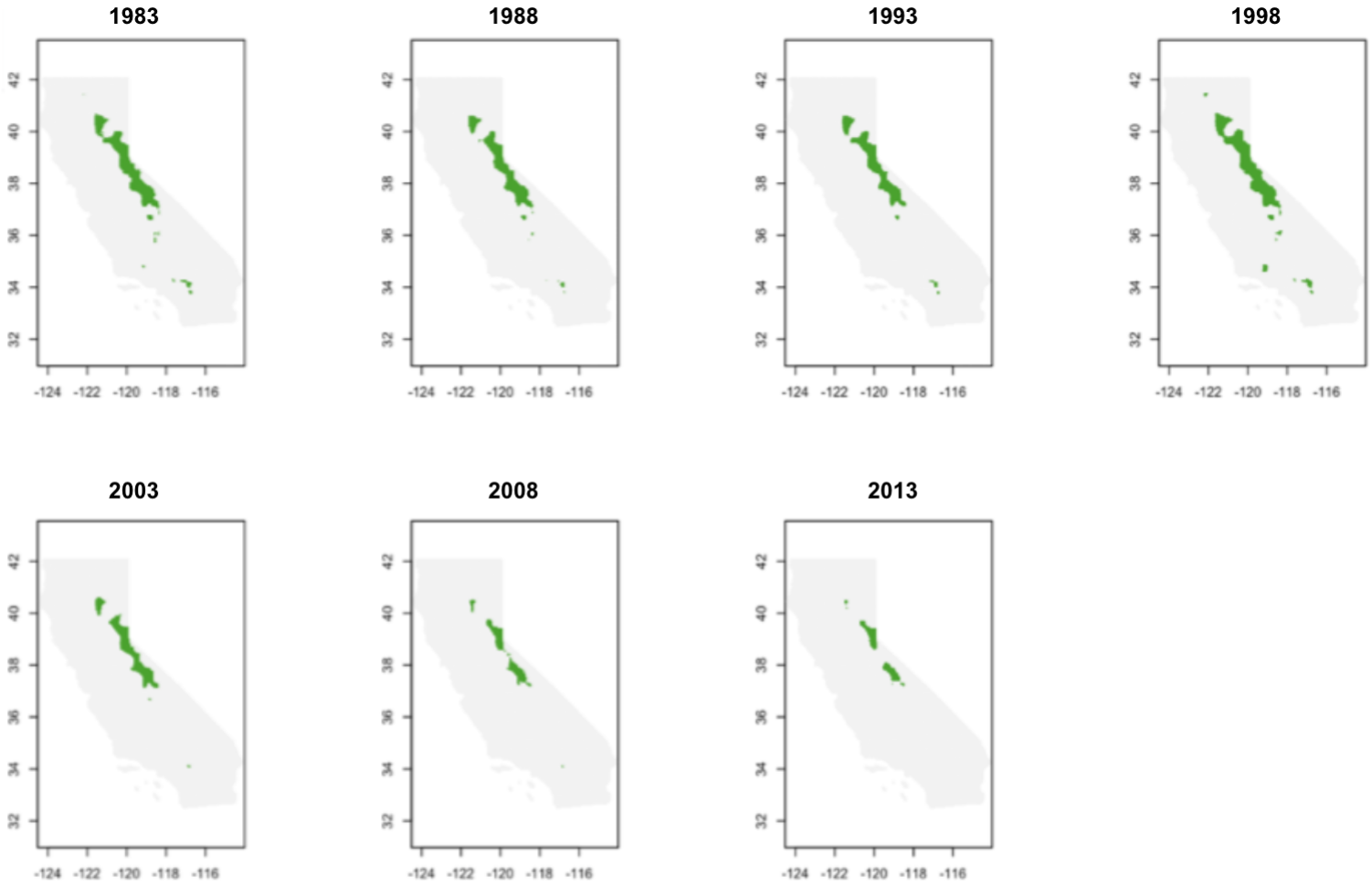

Figure 3.13: Plot of the areas (colored green) in which the posterior probability of the risk of plague exceeding 0.05 is greater than 0.95, for each time interval in the study.

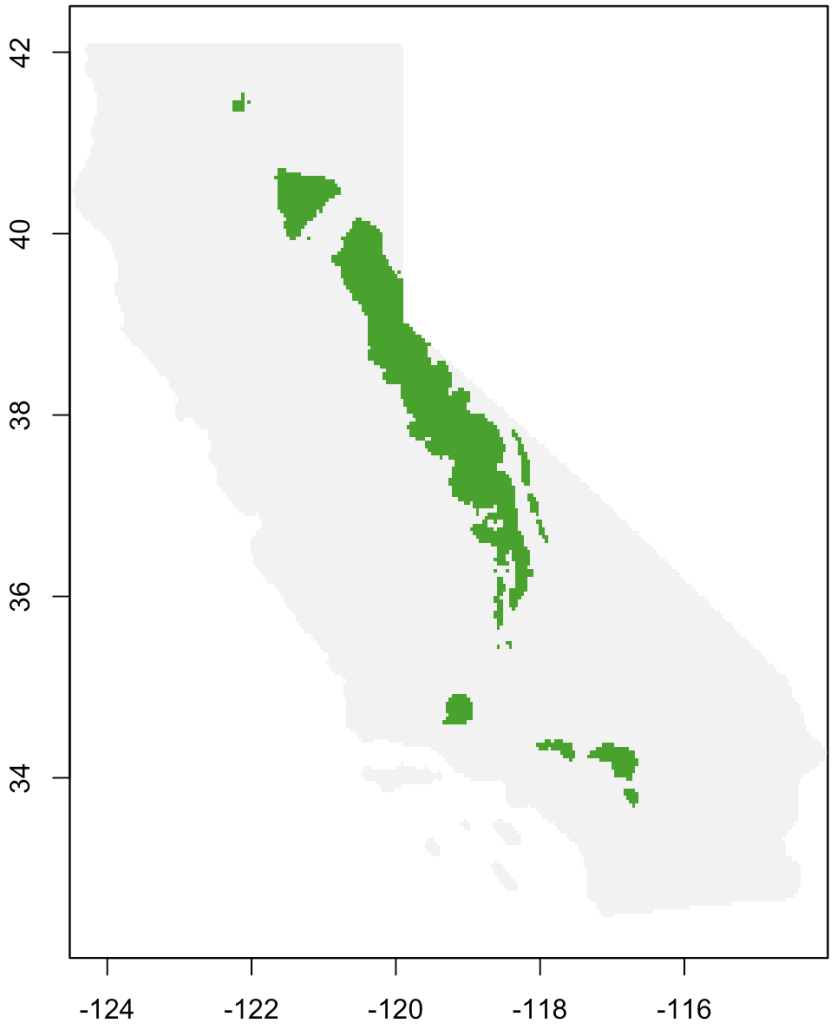

Figure 3.14: Plot of the areas (colored green) in which the posterior probability of the risk of plague exceeding 0.05 is greater than 0.95 over the aggregated timespan, between 1983 and 2015.

1983	1988	1993	1998	2003	2008	2013
0.702	0.554	0.556	0.802	0.459	0.262	0.199

Table 3.10: Fractional risk significance of the temporal model compared to aggregate model. Each column header identifies a time interval, while column values represent the fraction of the total number of pixels deemed significant in the temporal model divided by the total number of pixels deemed significant by the time-aggregated model. E.g., only 80.2% of the pixels calculated to be significant by the aggregate model were also deemed significant by the spatiotemporal model.

3.4.4 Discussion

In this analysis the proposed model captured important temporal trends overlooked in our previous analyses (Projects 1 and 2), as evidenced by differences in predicted risk between values of the proposed model and those of the reference, time-aggregated model. Both per-cell examinations of these differences (Figure 3.12) as well as a comparison of the confidence regions identifying areas of above average risk corroborate this conclusion (Figure 3.13, 3.14, Table 3.10). However, a number of remaining questions present themselves to the body of this discussion, namely what underlying factors explain the divergence between the temporal and aggregate models, what are the practical implications of the significance maps for sampling strategy, and perhaps most importantly, what are the limitations of the spatiotemporal structure present in the proposed model. We now address each of these questions in turn.

Regarding the nature of the difference in predicted risk between the temporal and aggregate models, we first characterize its direction and trend over time. For this purpose the scatterplots comparing cell-by-cell estimated risks are conducive. From Figure (3.12) the apparent trend is for the spatiotemporal method to yield generally lower risk estimates than the time-aggregated model. Furthermore, this difference tends to grow larger over time, especially for time intervals centered at or after 2003. This tendency can ultimately be explained by changes in the volume of data observed over time. Due to a moderate but consistent decrease in the number of observation

sites and collected specimen beginning in 2003, the earlier time intervals falling prior to 2003 exert a greater influence on the estimation of the spatial random field w in model (1.1). Simply put, if the true spatiotemporal process changes over time, the estimated random effects w, assumed constant over time in model (1.1), become biased towards those intervals possessing the most observations, which fall earlier in time. A similar phenomenon was observed in the simulation chapter of this project, which witnessed the aggregated model suffering worse root mean squared error in estimated log disease odds as time increased when the underlying temporal trend in risk was decreasing, and vice versa for an increasing temporal trend.

However the question remains whether the apparent decrease in estimated risk over time relates to a true decrease in the risk of the underlying disease, or is merely a result of attenuated sampling effort over time. To answer that question we must first address the interpretation of our estimated significance regions, presented in Figures (3.13) and (3.14). These plots encapsulate our spatially and temporally varying confidence that the risk of plague in any given pixel exceeds 0.05. Specifically, areas in which the posterior probability of risk being greater than 0.05 exceeds 0.95 are colored green. Since the overall prevalence of plague in Sciurids between 1983 and 2015 is roughly 0.05, we can conventiently interpret these maps to be areas in which we are highly confident that the risk of plague surpasses the average prevalence. We emphasize under this interpretation that areas deemed "significant" in the above maps need not necessarily be high risk in any absolute sense, for instance, possessing a risk of truly high magnitude, such as 0.12. Rather such areas are merely regions in which we have a high degree of certainty that the risk of plague is not 0.05 or below. Conversely, areas deemed "not significant" are not necessarily plague free. In reality, such areas could even possess substantial risk for plague in the Sciurid population, but, hypothetically, may possibly not register on the significance map due to a high posterior variance in estimated risk, typically brought about by a small sample size. For this reason, to

better ascertain whether the decrease over time witnessed in Figure (3.13) is affected primarily by a decrease in the actual underlying disease risk, or is merely an artifact of dampened sampling effort, it is helpful to consider an additional type of map depicting areas in which we are confident that the risk of plague is *less than* a particular value, here taken to be 0.05. To that end, Figures (3.15A, B) identify areas in which the probability that the risk of plague falls below 0.05 exceeds 0.95 for time the time intervals centered at 1983 and 2013, respectively. We can easily observe an increase in the coverage of the study region for which this condition holds true, indicating that our observed decrease in estimated disease risk is not merely due to a decrease in sampling effort alone, but is, to some extent, reflected in an underlying temporal change in risk. However, this conclusion does not imply that sampling effort has no impact on the observed decrease in areas deemed to be of significant (> 0.05) risk. On the contrary, to see where sampling effort bears particular weight, we turn to our next point, one which also can be used to inform sampling design.

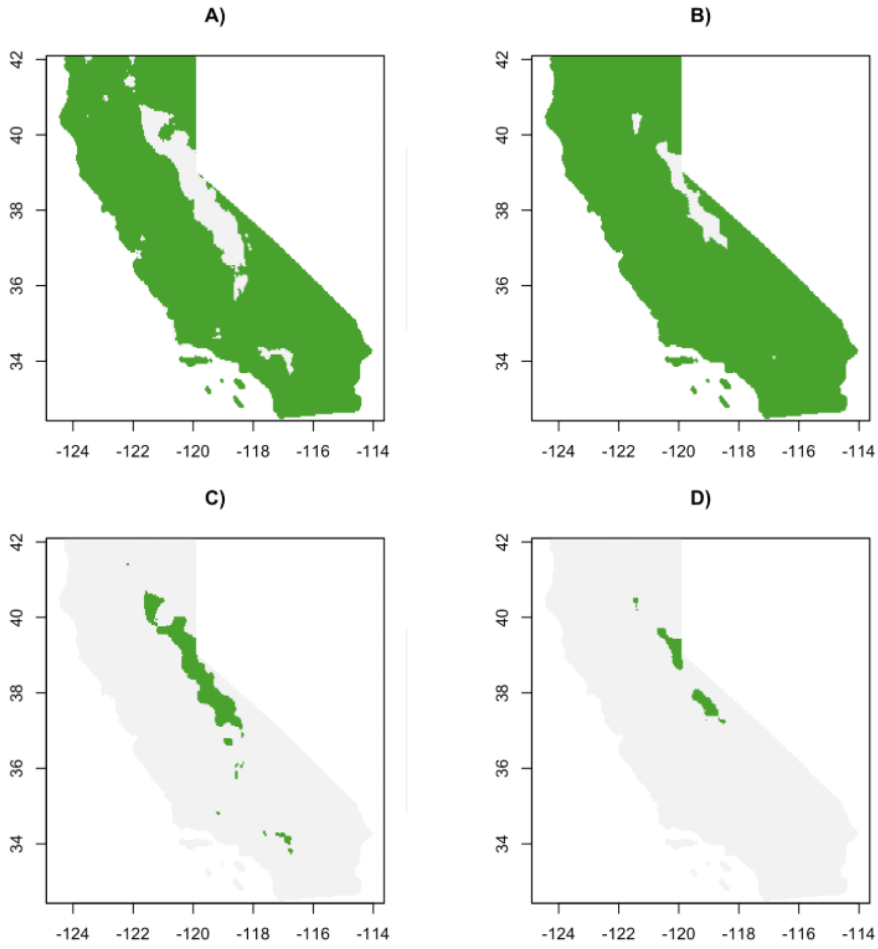

Figure 3.15: Positive and negative confidence regions for the risk of plague. In the top row, places where the probability that the risk of plague falls below 0.05 exceeds 0.95 in A) 1983 and B) 2013 are colored green. In the bottom row, green regions identify places where the probability of plague risk exceeding 0.05 is greater than 0.95 in C) 1983 and D) 2013.

An additional benefit of these maps showing areas of confidence for the risk of plague falling below a certain value, hereafter referred to as negative confidence regions, is that they can offer practical suggestions for future sampling effort, when viewed in conjunction with those maps showing areas of confidence for plague risk falling above that same value, which we denote positive confidence regions. For instance, the negative confidence regions of Figure (3.15B) identify a continuous band of space between the 40th and 36th parallel in which we are not confident that the risk of plague is below 0.05. But the positive confidence regions of Figure (3.15D) show this band overlapped by two regions in which we are confident that risk exceeds 0.05. The negative intersection of these regions of interest from Figure (3.15B) and Figure (3.15D), a narrow strip extending just above and below the 38th parallel (Figure 3.16), calls for special attention. As this region is such that we are neither confident that the risk of plague falls above or below 0.05, it may warrant further sampling effort in the future. This strategy of identifying areas in which we are not highly confident that risk exceeds or falls below the average value may be used to inform sampling design in general.

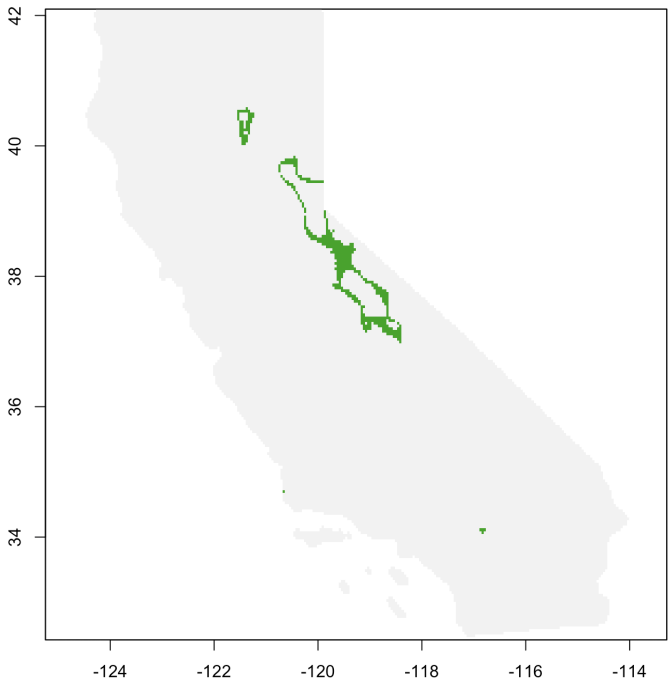

Figure 3.16: Areas suggested for additional sampling. Green pixels identify areas in which the probability of risk exceeding 0.05 is not above 0.95 and the probability of risk falling below 0.05 is not above 0.95, for the time interval centered at 2013.

Before discussing general modeling limitations we must present a strong caveat with regard to the above significance maps, one that relates to the drawbacks inherent in the concept of a P value. In frequentist statistics, particularly hypothesis testing, a P value is defined as the probability of observing a future test statistic more extreme than that currently observed if the null hypothesis is true. Specifically, $P = Pr(S(y) > S(y_0)|H)$, where $S(y)$ is the value of a test statistic under any future replication yielding data y and $S(y_0)$ is the value of the statistic given the observed data y_0. Under typical hypothesis testing the null hypothesis is rejected if $P < 0.05$. Criticisms of this use of P values abound (e.g., Gelman 2013), most notably, for our purpose here, the fact that the imposition of the arbitrary 0.05 cutoff does not discriminate between instances when P = 0.051 and when P = 0.049, which can lead to misleading conclusions. Our calculations of the above significance maps suffer from

this problematic imposition of an arbitrary cutoff, doubly so, first through the 0.95 posterior probability cutoff meant to identify our confidence, and secondly in the underlying 0.05 cutoff in estimated disease risk which forms the basis of our probability statement. Consequently, our maps could look very different depending on what posterior probability cutoff and what underlying risk cutoff are used.

To mitigate the impact of the arbitrary imposition of the 0.95 posterior probability threshold we can present significance maps with an extended color wheel. For instance, Figure (3.17) divides regions into 4 color categories depending on the magnitude of their posterior probabilities. Areas where these probabilities exceed 0.25, 0.5, and 0.95 are clearly identified, thus conveying a broader range of confidence levels than that offered by a simple 0.95 cutoff. While this solution is not perfect, our posterior probability and risk cutoffs ultimately can yield suggestions of practical importance (ex. Figure 3.16), and in general offer results consistent with the domain expertise of the wildlife biologists who operate the surveillance system.

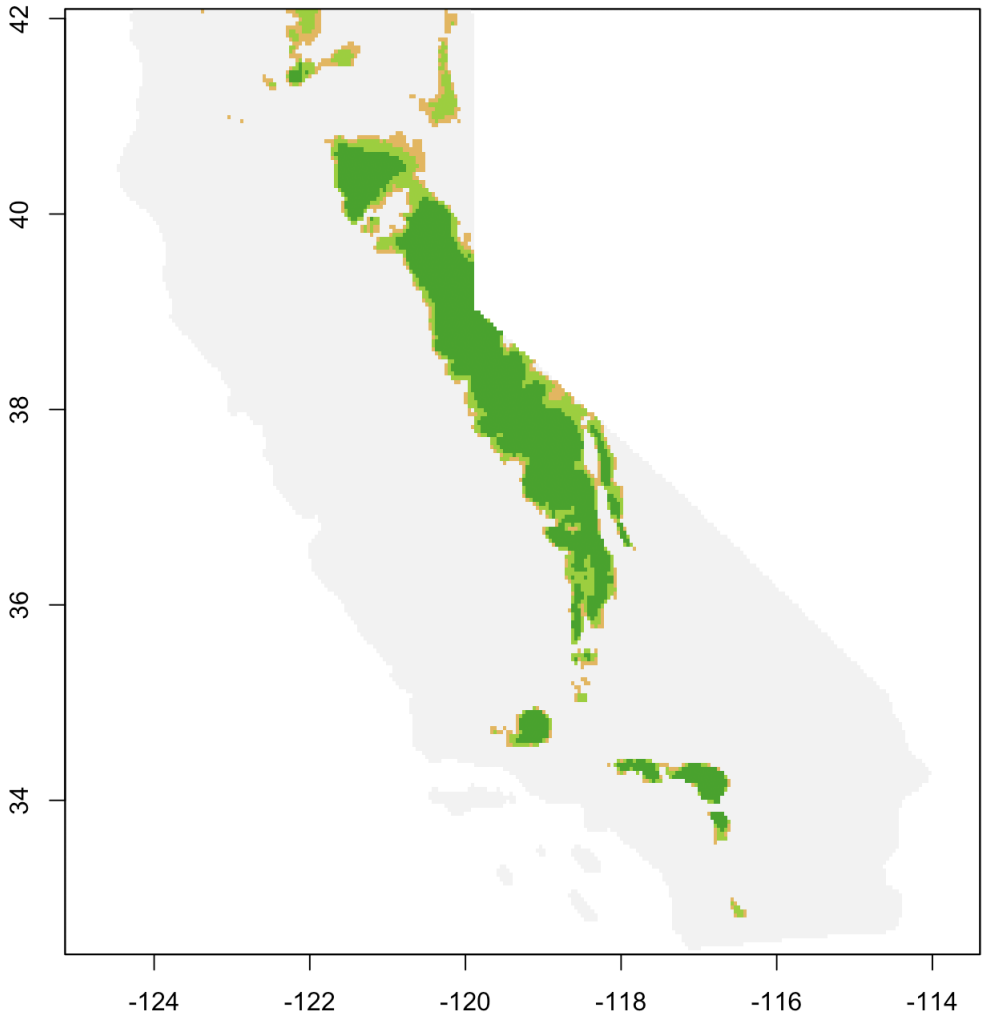

Figure 3.17: Areas where the posterior probability of plague risk exceeding 0.05 is greater than 0.95 (dark green), greater than 0.50 (light green) and 0.25 (brown) for Sciurids between 1983 and 2015.

Lastly, several simplifying assumptions behind the proposed model could be dispensed with to offer a more flexible spatiotemporal structure. Most prominent among these would be the transition from a discrete time to continuous time framework, under which the estimated disease risk would no longer be dependent upon the width of the temporal discretization window. In addition, more sophisticated spatial and temporal trends could be captured by incorporating spacetime interaction in the modeling structure, while the fixed effects β_m in model (3.1) could also be constructed to vary with time. Despite the limitations brought about by omitting these more complex features, this analysis has shown that the proposed model can nevertheless capture important temporal trends which are overlooked by the time-aggregated, benchmark approach, all the while remaining computationally tractable. We thus leave the aforementioned enhancements to future development.

Bibliography

[1] Ahn, Jaeil, et al. "A space-time point process model for analyzing and predicting case patterns of diarrheal disease in northwestern Ecuador." *Spatial and Spatio-temporal Epidemiology* 9 (2014): 23-35.

[2] Andrieu, Christophe, and Johannes Thoms. "A tutorial on adaptive MCMC." *Statistics and Computing* 18.4 (2008): 343-373.

[3] Banerjee, Sudipto, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical modeling and analysis for spatial data.* Chapman and Hall/CRC, 2014.

[4] Banerjee, Sudipto, Alan E. Gelfand, and Wolfgang Polasek. "Geostatistical modelling for spatial interaction data with application to postal service performance." *Journal of Statistical Planning and Inference* 90.1 (2000): 87-105.

[5] Banos, Arnaud, and Javier Lacasa. "Spatio-temporal exploration of SARS epidemic." *Cybergeo: European Journal of Geography* (2007).

[6] Benedict, Carol Ann. *Bubonic plague in nineteenth-century China.* Stanford University Press, 1996.

[7] Benes, Viktor, et al. "A case study on point process modelling in disease mapping." *Image Analysis & Stereology* 24.3 (2011): 159-168.

[8] Besag, Julian, Jeremy York, and Annie Mollie. "Bayesian image restoration,

with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics* 43.1 (1991): 1-20.

[9] Boyle, Phillip, and Marcus Frean. "Dependent gaussian processes." *Advances in Neural Information Processing Systems*. 2005.

[10] Bradley, Colleen A., et al. "BioSense: implementation of a national early event detection and situational awareness system." *MMWR Morb Mortal Wkly Rep* 54.Suppl (2005): 11-19.

[11] Brix, Anders, and Peter J. Diggle. "Spatiotemporal prediction for log-Gaussian Cox processes." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.4 (2001): 823-841.

[12] Brommer, Jon E. "The range margins of northern birds shift polewards." *Annales Zoologici Fennici.* Finnish Zoological and Botanical Publishing Board, 2004.

[13] Buller, Ian. On estimating the spatial distribution of enzootic *Yersinia pestis* in the United States using a wide-ranging sentinel species and spatial statistics with sampling considerations. 2019. Emory University, PhD dissertation.

[14] Byrne, Joseph Patrick. *The black death*. Greenwood Publishing Group, 2004.

[15] Cecconi, Lorenzo, et al. "Preferential sampling in veterinary parasitological surveillance." *Geospatial Health* (2016).

[16] Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander, J. A. (2011). "Point pattern modelling for degraded presence-only data over large regions." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5), 757-776.

[17] Chanteau, Suzanne, et al. "Early diagnosis of bubonic plague using F1 antigen capture ELISA assay and rapid immunogold dipstick." *International Journal of Medical Microbiology* 290.3 (2000): 279-283.

[18] Chen, Cheng, et al. "An improved spatial downscaling procedure for TRMM 3B43 precipitation product using geographically weighted regression." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.9 (2015): 4592-4604.

[19] Childs, James E., et al. "Predicting the local dynamics of epizootic rabies among raccoons in the United States." *Proceedings of the National Academy of Sciences* 97.25 (2000): 13666-13671.

[20] Chiles, Jean-Paul, and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*. Vol. 497. John Wiley & Sons, 2009.

[21] Choi, Jihye, et al. "Web-based infectious disease surveillance systems and public health perspectives: a systematic review." *BMC Public Health* 16.1 (2016): 1238.

[22] Chu, Wei, and Zoubin Ghahramani. "Gaussian processes for ordinal regression." *Journal of Machine Learning Research* 6.Jul (2005): 1019-1041.

[23] Conn, Paul B., James T. Thorson, and Devin S. Johnson. "Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage." *Methods in Ecology and Evolution* 8.11 (2017): 1535-1546.

[24] Cox, D. R. "Some statistical methods related with series of event." *J. Roy. Soc. B.* Vol. 17. No. 2. 1955.

[25] Diggle, Peter, Barry Rowlingson, and Ting-li Su. "Point process methodology

for online spatio-temporal disease surveillance." *Environmetrics: The Official Journal of the International Environmetrics Society* 16.5 (2005): 423-434.

[26] Diggle, Peter J. "Spatio-temporal point processes, partial likelihood, foot and mouth disease." *Statistical Methods in Medical Research* 15.4 (2006): 325-336.

[27] Diggle, Peter J., Raquel Menezes, and Tingli Su. "Geostatistical inference under preferential sampling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2 (2010): 191-232.

[28] Diggle, Peter J. *Statistical analysis of spatial and spatio-temporal point patterns.* Chapman and Hall/CRC, 2013.

[29] Dinsdale, Daniel, and Matias Salibian-Barrera. "Methods for preferential sampling in geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.1 (2019): 181-198.

[30] Dorazio, R. M. (2012). "Predicting the geographic distribution of a species from presence-only data subject to detection errors." *Biometrics*, 68(4), 1303-1312.

[31] Dorazio, R. M. (2014). "Accounting for imperfect detection and survey bias in statistical analysis of presence-only data." *Global Ecology and Biogeography*, 23(12), 1472-1484.

[32] Duane, Simon, et al. "Hybrid monte carlo." *Physics Letters B* 195.2 (1987): 216-222.

[33] Dubrule, Olivier. "Comparing splines and kriging." *Computers & Geosciences* 10.2-3 (1984): 327-338.

[34] Eisen, Rebecca J., et al. "Early-phase transmission of Yersinia pestis by unblocked fleas as a mechanism explaining rapidly spreading plague epizootics." *Proceedings of the National Academy of Sciences* 103.42 (2006): 15380-15385.

[35] Elith, Jane, and John R. Leathwick. "Species distribution models: ecological explanation and prediction across space and time." *Annual Review of Ecology, Evolution, and Systematics* 40 (2009): 677-697.

[36] Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). "Bias correction in species distribution models: pooling survey and collection data for multiple species." *Methods in Ecology and Evolution*, 6(4), 424-438.

[37] Fink, Daniel, et al. "Spatiotemporal exploratory models for broad-scale survey data." *Ecological Applications* 20.8 (2010): 2131-2147.

[38] Fink, Daniel, et al. "Crowdsourcing meets ecology: hemisphere-wide spatiotemporal species distribution models." *AI Magazine* 35.2 (2014): 19-30.

[39] Fletcher, R. J., McCleery, R. A., Greene, D. U., & Tye, C. A. (2016). "Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions." *Landscape Ecology*, 31(6), 1369-1382.

[40] Furrer, Reinhard, Marc G. Genton, and Douglas Nychka. "Covariance tapering for interpolation of large spatial datasets." *Journal of Computational and Graphical Statistics* 15.3 (2006): 502-523.

[41] Gelfand, Alan E., et al. "Nonstationary multivariate process modeling through spatially varying coregionalization." Test 13.2 (2004): 263-312.

[42] Gelfand, Alan E., Sujit K. Sahu, and David M. Holland. "On the effect of preferential sampling in spatial prediction." *Environmetrics* 23.7 (2012): 565-578.

[43] Gelfand, Alan E., and Shinichiro Shirota. "Preferential sampling for pres-

ence/absence data and for fusion of presence/absence data with presence-only data." *Ecological Monographs* 89.3 (2019): e01372.

[44] Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2016). "Capitalizing on opportunistic data for monitoring relative abundances of species." *Biometrics*, 72(2), 649-658.

[45] Girolami, Mark, and Simon Rogers. "Variational Bayesian multinomial probit regression with Gaussian process priors." *Neural Computation* 18.8 (2006): 1790-1817.

[46] Goulard, Michel, and Marc Voltz. "Linear coregionalization model: tools for estimation and choice of cross-variogram matrix." *Mathematical Geology* 24.3 (1992): 269-286.

[47] Grzebyk, Michel, and Hans Wackernagel. "Multivariate analysis and spatial/temporal scales: real and complex models." *Proceedings of the XVIIth International Biometrics Conference*. Vol. 1. Hamilton Ontario, 1994.

[48] Hanks, Ephraim M., Mevin B. Hooten, and Fred A. Baker. "Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence." *Ecological Applications* 21.4 (2011): 1173-1188.

[49] Hefley, T. J., & Hooten, M. B. (2016). "Hierarchical species distribution models." *Current Landscape Ecology Reports*, 1(2), 87-97.

[50] Helser, Thomas E., Andr E. Punt, and Richard D. Methot. "A generalized linear mixed model analysis of a multi-vessel fishery resource survey." *Fisheries Research* 70.2-3 (2004): 251-264.

[51] Higdon, David. "A process-convolution approach to modelling temperatures in

the North Atlantic Ocean." *Environmental and Ecological Statistics* 5.2 (1998): 173-190.

[52] Higdon, Dave. "Space and space-time modeling using process convolutions." *Quantitative Methods for Current Environmental Issues*. Springer, London, 2002. 37-56.

[53] Hoffman, Matthew D., and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15.1 (2014): 1593-1623.

[54] Hooten, Mevin B., and Christopher K. Wikle. "Shifts in the spatio-temporal growth dynamics of shortleaf pine." *Environmental and Ecological Statistics* 14.3 (2007): 207-227.

[55] Horn, R. A., C. R. Johnson, and L. Elsner. "Topics and Matrix Analysis." *Jahresbericht der Deutschen Mathematiker Vereinigung* 96.4 (1994): 79-79.

[56] Johnson, Glen D. "Prospective spatial prediction of infectious disease: experience of New York State (USA) with West Nile Virus and proposed directions for improved surveillance." *Environmental and Ecological Statistics* 15.3 (2008): 293-311.

[57] Journel, Andre G., and Charles J. Huijbregts. Mining geostatistics. Vol. 600. London: Academic press, 1978.

[58] Kim, Hyun-Chul, and Zoubin Ghahramani. "Bayesian Gaussian process classification with the EM-EP algorithm." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.12 (2006): 1948-1959.

[59] Kleinman, Ken, Ross Lazarus, and Richard Platt. "A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an

application to biological terrorism." *American Journal of Epidemiology* 159.3 (2004): 217-224.

[60] Knorr-Held, Leonhard, and Julian Besag. "Modelling risk from a disease in time and space." *Statistics in Medicine* 17.18 (1998): 2045-2060.

[61] Knorr-Held, Leonhard. "Bayesian modelling of inseparable space-time variation in disease risk." *Statistics in Medicine* 19.17-18 (2000): 2555-2567.

[62] Knox, E. G., and M. S. Bartlett. "The detection of space-time interactions." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 13.1 (1964): 25-30.

[63] Kottas, Athanasios, Jason A. Duan, and Alan E. Gelfand. "Modeling disease incidence data with spatial and spatio temporal Dirichlet process mixtures." *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.1 (2008): 29-42.

[64] Kulldorff, Martin, and Neville Nagarwalla. "Spatial disease clusters: detection and inference." *Statistics in Medicine* 14.8 (1995): 799-810.

[65] Kulldorff, Martin. "Prospective time periodic geographical disease surveillance using a scan statistic." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164.1 (2001): 61-72.

[66] Kustas, William P., et al. "Estimating subpixel surface temperatures and energy fluxes from the vegetation index-radiometric temperature relationship." *Remote Sensing of Environment* 85.4 (2003): 429-440.

[67] Latimer, Andrew M., et al. "Building statistical models to analyze species distributions." *Ecological Applications* 16.1 (2006): 33-50.

[68] Lee, Duncan, Claire Ferguson, and E. Marian Scott. "Constructing representative air quality indicators with measures of uncertainty." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174.1 (2011): 109-126.

[69] Lee, Duncan, and Richard Mitchell. "Boundary detection in disease mapping studies." *Biostatistics* 13.3 (2012): 415-426.

[70] Lee, Duncan, and Christophe Sarran. "Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies." *Environmetrics* 26.7 (2015): 477-487.

[71] Lee, A., et al. "Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology." *Environmetrics* 26.4 (2015): 255-267.

[72] Leroux, Brian G., Xingye Lei, and Norman Breslow. "Estimation of disease rates in small areas: a new mixed model for spatial dependence." *Statistical Models in Epidemiology, the Environment, and Clinical Trials.* Springer, New York, NY, 2000. 179-191.

[73] Li, Qi, and Jeff Racine. "Cross-validated local linear nonparametric regression." *Statistica Sinica* 14.2 (2004): 485-512.

[74] Li, Yehua, and Yongtao Guan. "Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance." *Journal of the American Statistical Association* 109.507 (2014): 1205-1215.

[75] Liang, Shengde, Bradley P. Carlin, and Alan E. Gelfand. "Analysis of Minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information." *The Annals of Applied Statistics* 3.3 (2008): 943.

[76] Lin, Xiwu, et al. "Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV." *The Annals of Statistics* 28.6 (2000): 1570-1600.

[77] Link, William A., and John R. Sauer. "Seasonal components of avian population change: joint analysis of two large-scale monitoring programs." Ecology 88.1 (2007): 49-55.

[78] Maclean, Ilya MD, et al. "Climate change causes rapid changes in the distribution and site abundance of birds in winter." *Global Change Biology* 14.11 (2008): 2489-2500.

[79] Maggini, Ramona, et al. "Improving generalized regression analysis for the spatial prediction of forest communities." *Journal of Biogeography* 33.10 (2006): 1729-1749.

[80] Merlin, Olivier, et al. "A sequential model for disaggregating near-surface soil moisture observations using multi-resolution thermal sensors." *Remote Sensing of Environment* 113.10 (2009): 2275-2284.

[81] Michalcova, D., Lvonck, S., Chytry, M., & Hajek, O. (2011). "Bias in vegetation databases? A comparison of stratified-random and preferential sampling." *Journal of Vegetation Science*, 22(2), 281-291.

[82] Migliani, Rene, et al. "Epidemiological trends for human plague in Madagascar during the second half of the 20th century: a survey of 20 900 notified cases." *Tropical Medicine & International Health* 11.8 (2006): 1228-1237.

[83] Mitchell, Peter J., Jacquomo Monk, and Laurie Laurenson. "Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes." *Methods in Ecology and Evolution* 8.1 (2017): 12-21.

[84] Naus, Joseph I. "The distribution of the size of the maximum cluster of points on a line." *Journal of the American Statistical Association* 60.310 (1965): 532-538.

[85] Neal, Radford M. "MCMC using Hamiltonian dynamics." *Handbook of Markov Chain Monte Carlo* 2.11 (2011): 2.

[86] Nesterov, Yurii. "Primal-dual subgradient methods for convex problems." *Mathematical Programming* 120.1 (2009): 221-259.

[87] Nychka, Douglas, et al. "A multiresolution Gaussian process model for the analysis of large spatial datasets." *Journal of Computational and Graphical Statistics* 24.2 (2015): 579-599.

[88] Openshaw, Stan, et al. "A mark 1 geographical analysis machine for the automated analysis of point data sets." *International Journal of Geographical Information System* 1.4 (1987): 335-358.

[89] Paci, Lucia, et al. "Spatial hedonic modelling adjusted for preferential sampling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2019).

[90] Pacifici, Krishna, et al. "Occupancy estimation for rare species using a spatially-adaptive sampling design." *Methods in Ecology and Evolution* 7.3 (2016): 285-293.

[91] Pati, Debdeep, Brian J. Reich, and David B. Dunson. "Bayesian Geostatistical Modelling with Informative Sampling Locations." *Biometrika* 98.1 (2011): 35-48.

[92] Pennington, Michael. "Efficient estimators of abundance, for fish and plankton surveys." *Biometrics* (1983): 281-286.

[93] Pennino, Maria Grazia, et al. "Accounting for preferential sampling in species distribution models." *Ecology and Evolution* 9.1 (2019): 653-663.

[94] Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. "Maximum entropy modeling of species geographic distributions." *Ecological Modelling* 190.3-4 (2006): 231-259.

[95] Plowright, Raina K., et al. "Sampling to elucidate the dynamics of infections in reservoir hosts." *Philosophical Transactions of the Royal Society* B 374.1782 (2019): 20180336.

[96] Quick, Harrison, Lance A. Waller, and Michele Casper. "A multivariate space-time model for analysing county level heart disease death rates by race and sex." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.1 (2018): 291-304.

[97] Reich, Brian J., Howard H. Chang, and Kristen M. Foley. "A spectral method for spatial downscaling." *Biometrics* 70.4 (2014): 932-942.

[98] Reis, Ben Y., et al. "AEGIS: a robust and scalable real-time public health surveillance system." *Journal of the American Medical Informatics Association* 14.5 (2007): 581-588.

[99] Renner, Ian W., et al. "Point process models for presence-only analysis." *Methods in Ecology and Evolution* 6.4 (2015): 366-379.

[100] Rinaldi, Laura, et al. "Sheep and Fasciola hepatica in Europe: the GLOWORM experience." *Geospatial Health* (2015): 309-317.

[101] Rocchini, D., Garzon-Lopez, C. X., Marcantonio, M., Amici, V., Bacaro, G., Bastin, L., ... & He, K. S. (2017). "Anticipating species distributions: Han-

dling sampling effort bias under a Bayesian framework." *Science of the Total Environment*, 584, 282-290.

[102] Royle, J. A., & Link, W. A. (2006). "Generalized site occupancy models allowing for false positive and false negative errors." *Ecology*, 87(4), 835-841.

[103] Royle, J. A., Kery, M., Gautier, R., & Schmid, H. (2007). "Hierarchical spatial models of abundance and occurrence from imperfect survey data." *Ecological Monographs*, 77(3), 465-481.

[104] Robertson, Colin, et al. "Review of methods for space-time disease surveillance." *Spatial and Spatio-temporal Epidemiology* 1.2-3 (2010): 105-116.

[105] Royle, J. Andrew, et al. "Hierarchical spatial models of abundance and occurrence from imperfect survey data." *Ecological Monographs* 77.3 (2007): 465-481.

[106] Rue, H., et al. "Discussion on the paper Geostatistical inference under preferential sampling by Diggle, Menezes and Su." *Journal of the Royal Statistical Society*, Series C 52.221-223 (2010): 139-140.

[107] Shaddick, Gavin, and James V. Zidek. "A case study in preferential sampling: Long term monitoring of air pollution in the UK." *Spatial Statistics* 9 (2014): 51-65.

[108] Smith, D. L., et al. "Assessing the role of long-distance translocation and spatial heterogeneity in the raccoon rabies epidemic in Connecticut." *Preventive Veterinary Medicine* 71.3-4 (2005): 225-240.

[109] Stein, Michael L. "The screening effect in kriging." *The Annals of Statistics* 30.1 (2002): 298-323.

[110] Stephen, C., P. Zimmer, and M. Lee. "Is there a due diligence standard for

wildlife disease surveillance? A Canadian case study." *The Canadian Veterinary Journal= La Revue Veterinaire Canadienne* 60.8 (2019): 841.

[111] Seeger, Matthias, Yee-Whye Teh, and Michael Jordan. "Semiparametric latent factor models." *Artificial Intelligence and Statistics (AISTATS)*, 2005.

[112] Tabak, Michael A., Kerri Pedersen, and Ryan S. Miller. "Detection error influences both temporal seroprevalence predictions and risk factors associations in wildlife disease models." *Ecology and Evolution* (2019).

[113] Thomas, Chris D., and Jack J. Lennon. "Birds extend their ranges northwards." *Nature* 399.6733 (1999): 213.

[114] Thorson, James T., and Eric J. Ward. "Accounting for space-time interactions in index standardization models." *Fisheries Research* 147 (2013): 426-433.

[115] Thorson, James T., et al. "Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring." *Global Ecology and Biogeography* 25.9 (2016): 1144-1158.

[116] Veneziano, Daniele, and Peter K. Kitanidis. "Sequential sampling to contour an uncertain function." *Journal of the International Association for Mathematical Geology* 14.5 (1982): 387-404.

[117] Ver Hoef, Jay M., and Ronald Paul Barry. "Constructing and fitting models for cokriging and multivariable spatial prediction." *Journal of Statistical Planning and Inference* 69.2 (1998): 275-294.

[118] Wakefield, Jon, and Gavin Shaddick. "Health-exposure modeling and the ecological fallacy." *Biostatistics* 7.3 (2006): 438-455.

[119] Waller, Lance A., et al. "Hierarchical spatio-temporal mapping of disease rates." *Journal of the American Statistical Association* 92.438 (1997): 607-617.

[120] Waller, Lance A., and Carol A. Gotway. *Applied spatial statistics for public health data.* Vol. 368. John Wiley & Sons, 2004.

[121] Walther, Gian-Reto, et al. "Ecological responses to recent climate change." *Nature* 416.6879 (2002): 389.

[122] Ward, Eric J., et al. "Using spatiotemporal species distribution models to identify temporally evolving hotspots of species co-occurrence." *Ecological Applications* 25.8 (2015): 2198-2209.

[123] Wackernagel, Hans. "Multivariate geostatistics: an introduction with applications". *Springer Science & Business Media*, 2013.

[124] Williams, Christopher KI, and Carl Edward Rasmussen. "Gaussian processes for regression." *Advances in Neural Information Processing Systems*. 1996.

[125] Williams, Christopher KI, and David Barber. "Bayesian classification with Gaussian processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.12 (1998): 1342-1351.

[126] Wikle, Christopher K., and Noel Cressie. "A dimension-reduced approach to space-time Kalman filtering." *Biometrika* 86.4 (1999): 815-829.

[127] Yu, Hwa-Lung, Shang-Chen Ku, and Alexander Kolovos. "A GIS tool for spatiotemporal modeling under a knowledge synthesis framework." *Stochastic Environmental Research and Risk Assessment* 30.2 (2016): 665-679.